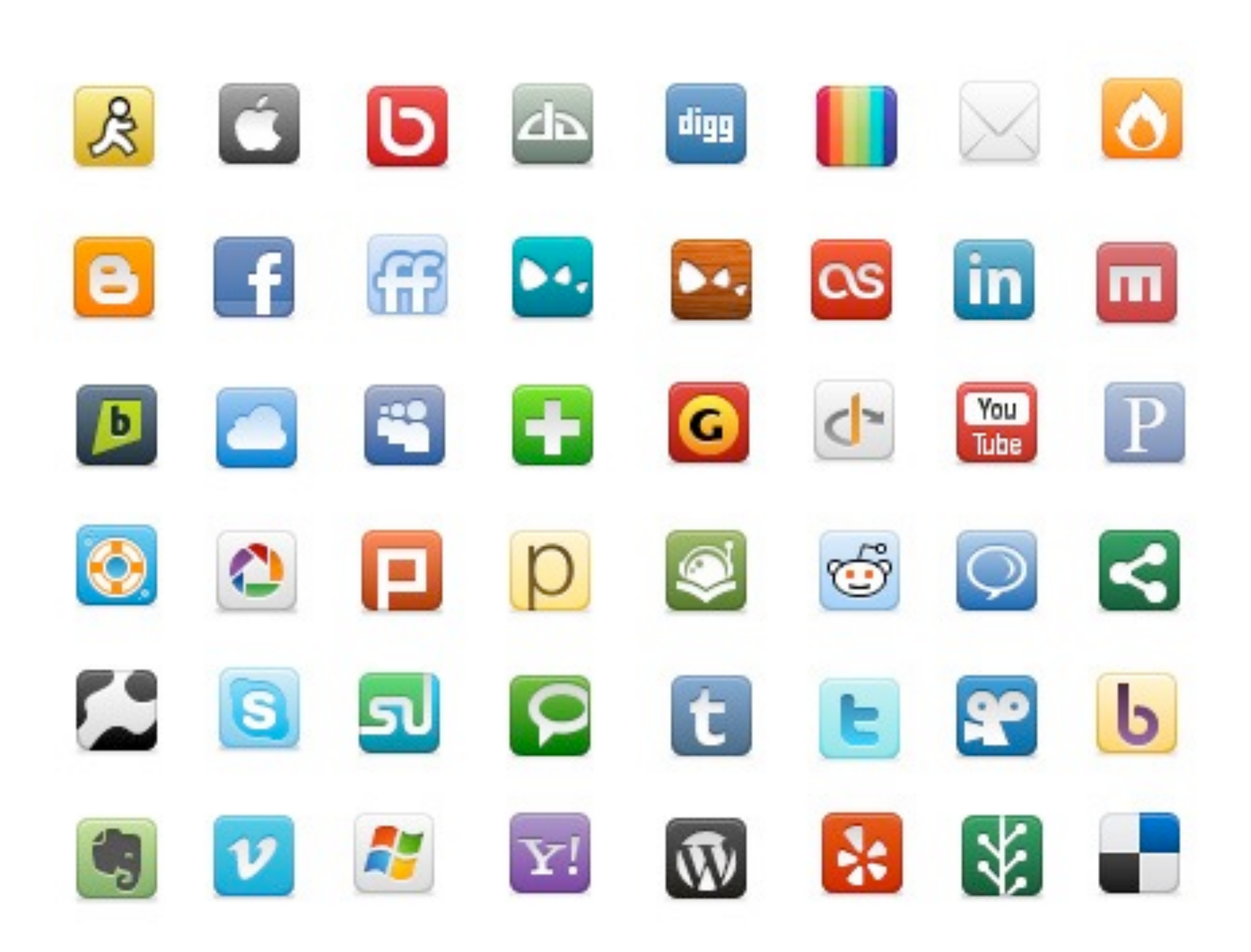
Analyse des réseaux sociaux et apprentissage

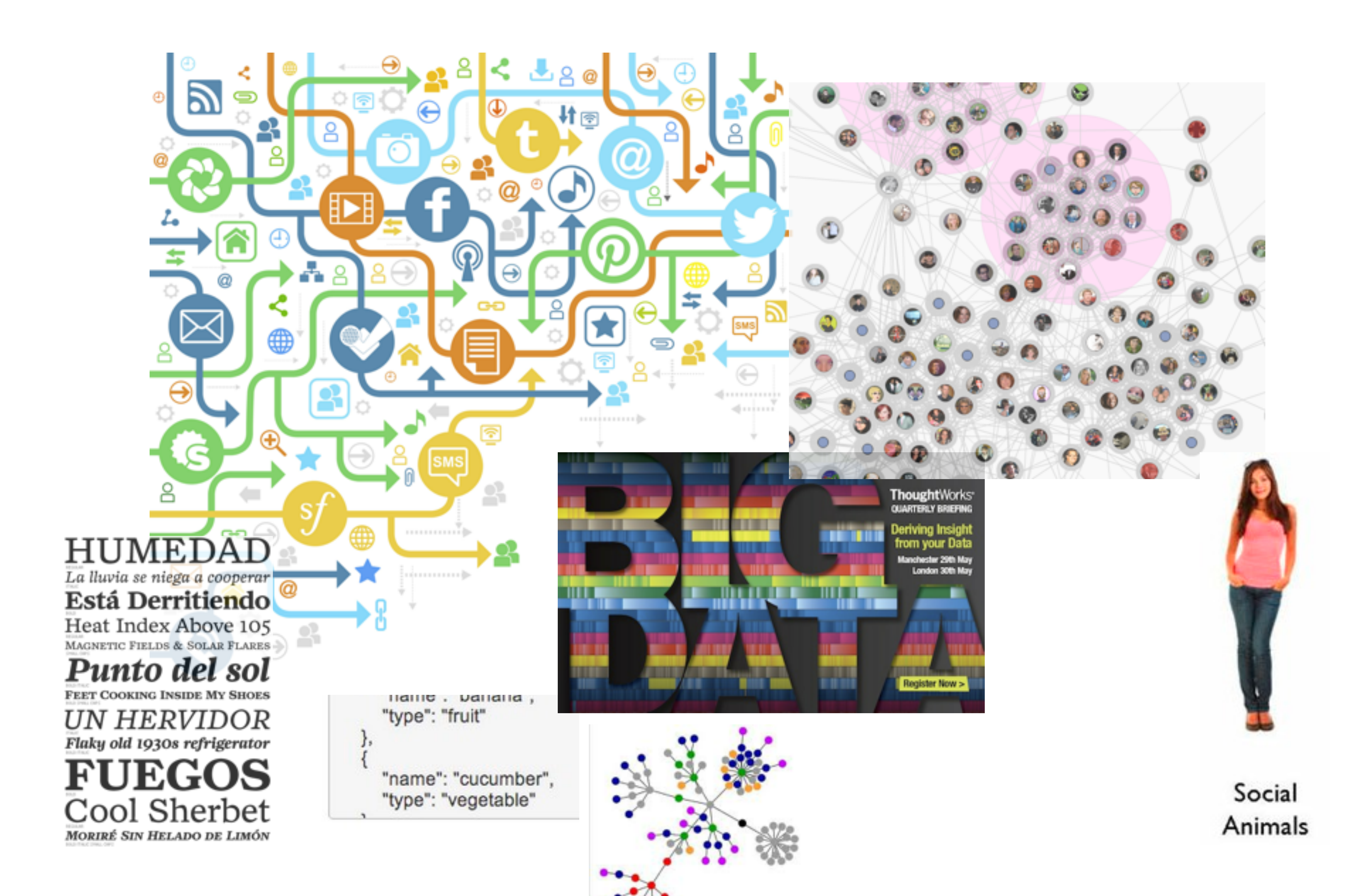
Emmanuel Viennet

Laboratoire de Traitement et Transport de l'Information
Université Paris 13 - Sorbonne Paris Cité

Réseaux sociaux ?



Réseaux sociaux ?



Analyse des réseaux sociaux ?

Fouille de données

Sociologie

Probabilités/Statistiques

Apprentissage

Théorie des graphes

Link Analysis

Algorithmique

Données textuelles (TAL)

Visualisation

Multimédia

Modélisation des réseaux sociaux

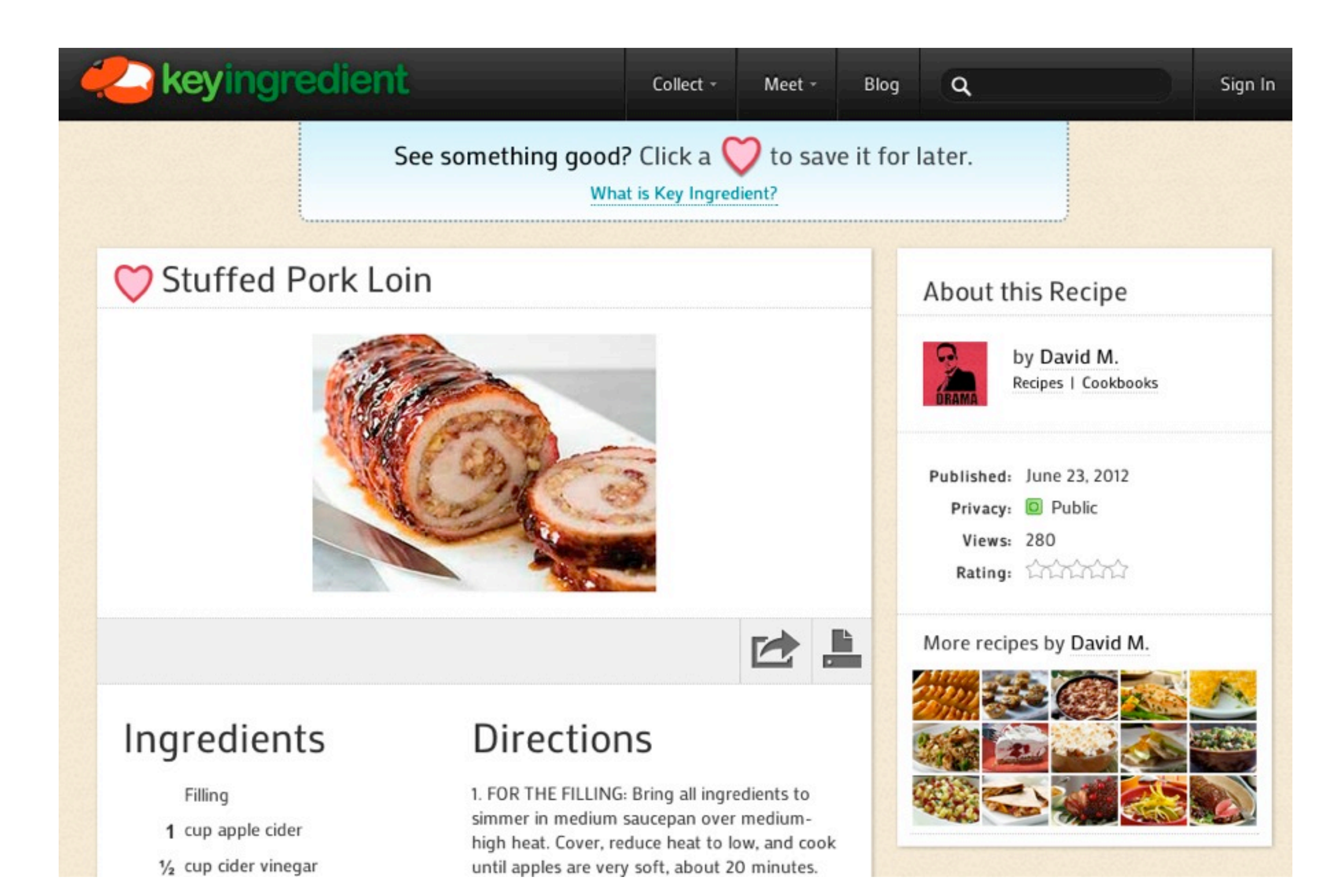
Modélisation par un *graphe*

- liens valués ou non, dirigés ou non
- nœuds porteurs d'attributs
- nœuds et liens dépendants du temps

Le graphe possède en général des propriétés structurelles particulières

Les variables observées ne sont pas iid

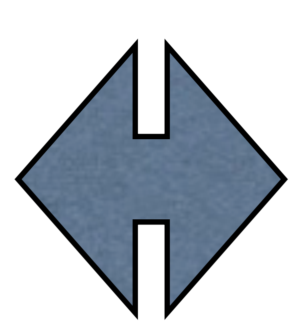
Contexte applicatif: un exemple



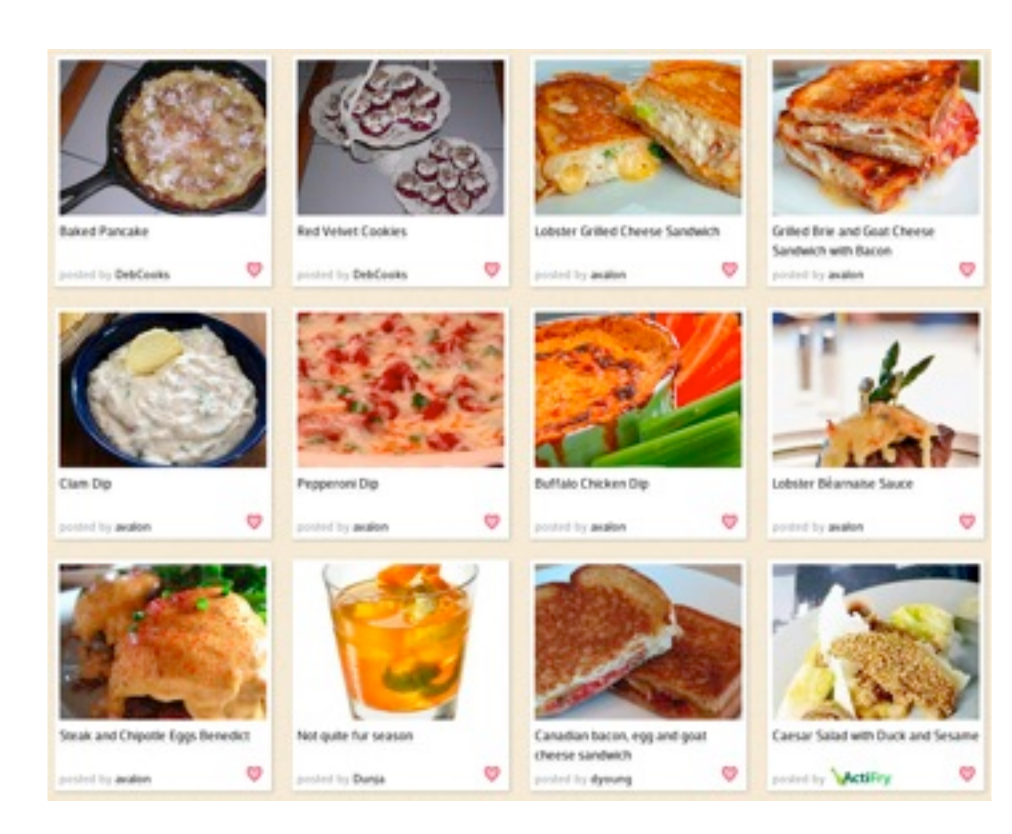
Contexte applicatif: un exemple

Utilisateurs





Recettes



- Blogs (de l'utilisateur ou de la recette)
- Notes
- Tags
- Commentaires sur les recettes
- Similarité entre recettes (texte, image, ingrédients)

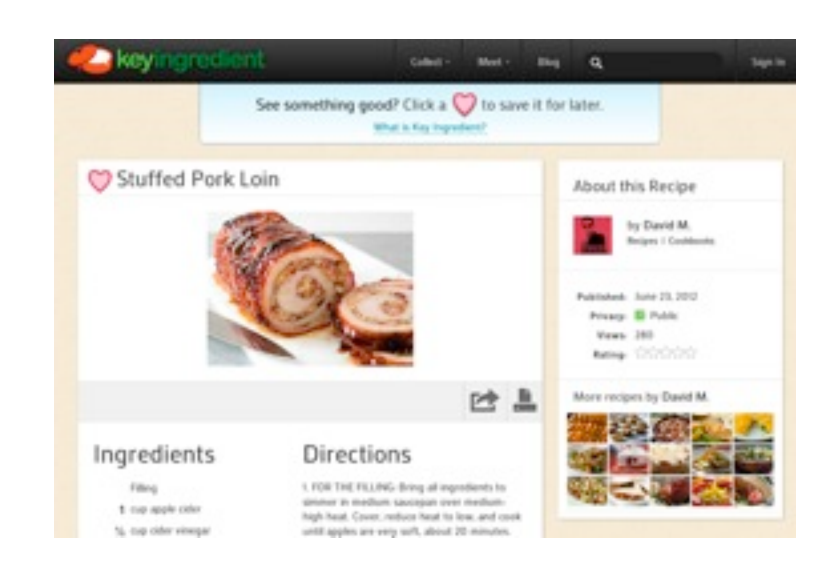
Contexte applicatif: un exemple

On peut construire de nombreux graphes...



Unipartis ou bipartis, «explicites» ou «implicites»

Quelques questions pertinentes:



- Analyser le comportement des utilisateurs
  - animation de communautés
    - ➡ groupes actifs / évolution
    - ➡ thèmes «chauds»
  - prévision de *churn*
- Recommander
  - suggérer des recettes à un utilisateur
  - suggérer des «amis»
- Modéliser diffusion et influence
  - viralité, épidémiologie, marketing
- Détection de fraude / identité

Exemples de données réelles traitées par notre équipe

Appels téléphoniques	10 -100 M
Blogs	10 M
Achats (e-commerce)	10 M
Sites web	1000

Parcimonie (*sparseness*): nombre de liens *proportionnel* au nombre de nœuds

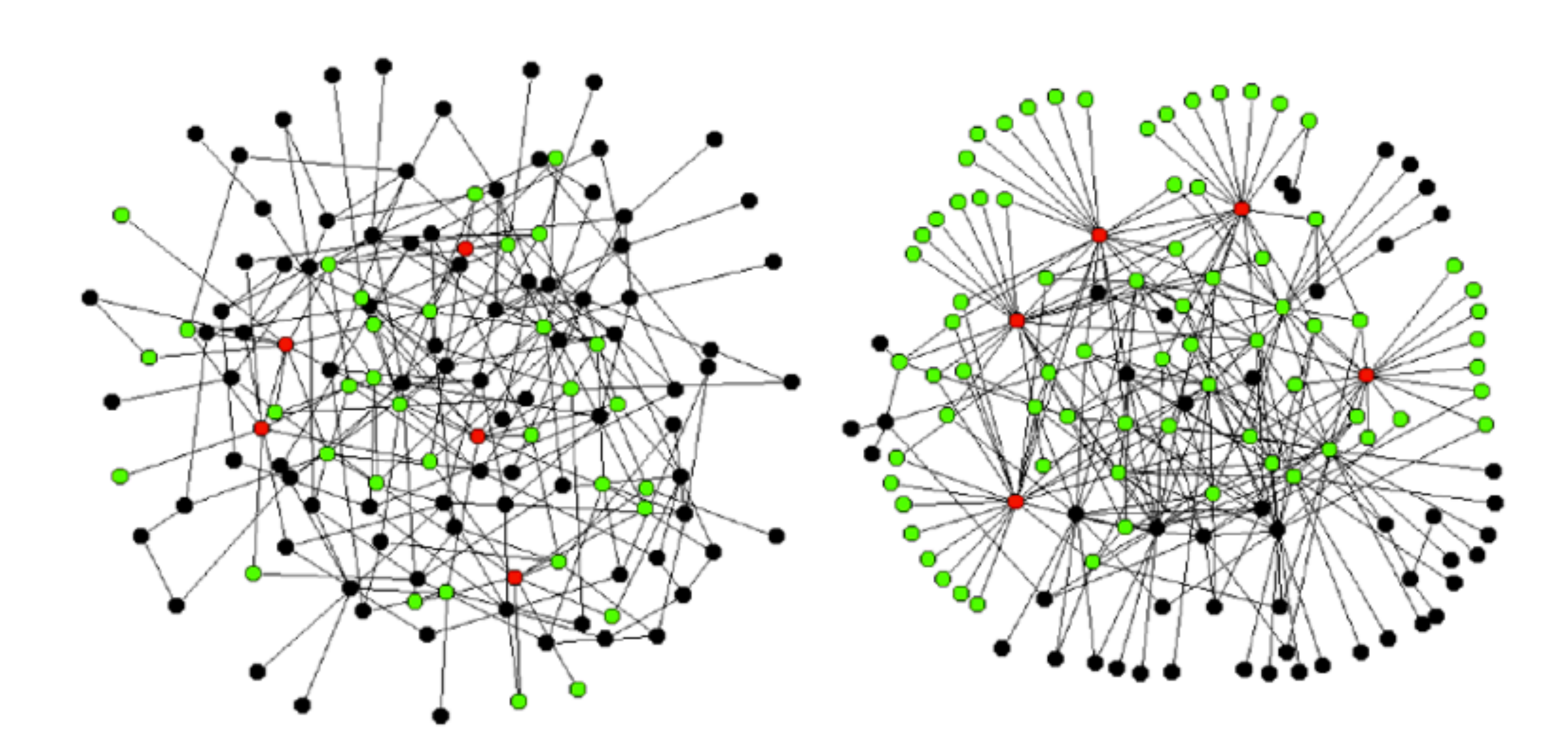
Graphes
(«complex networks»)

Exemple de propriété structurelle: l'effet petit monde

- Longueur moyenne du plus court chemin reliant deux noeuds petite

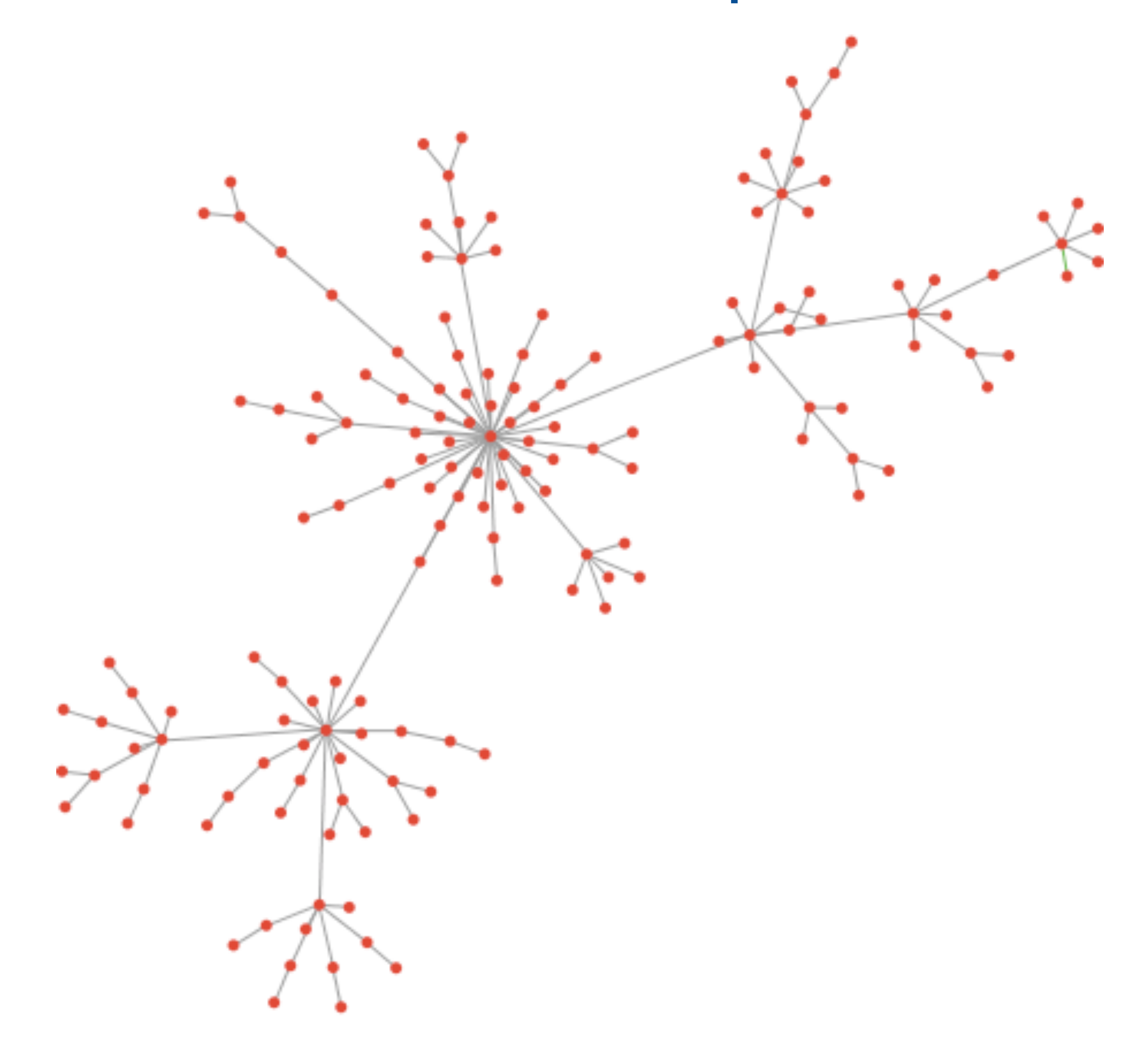
- Caractéristique liée à la distribution des degrés: graphe *scale free*

$$P(\text{degré d'un nœud } = k) \propto k^{-\gamma}$$



(Albert et al, 2000)

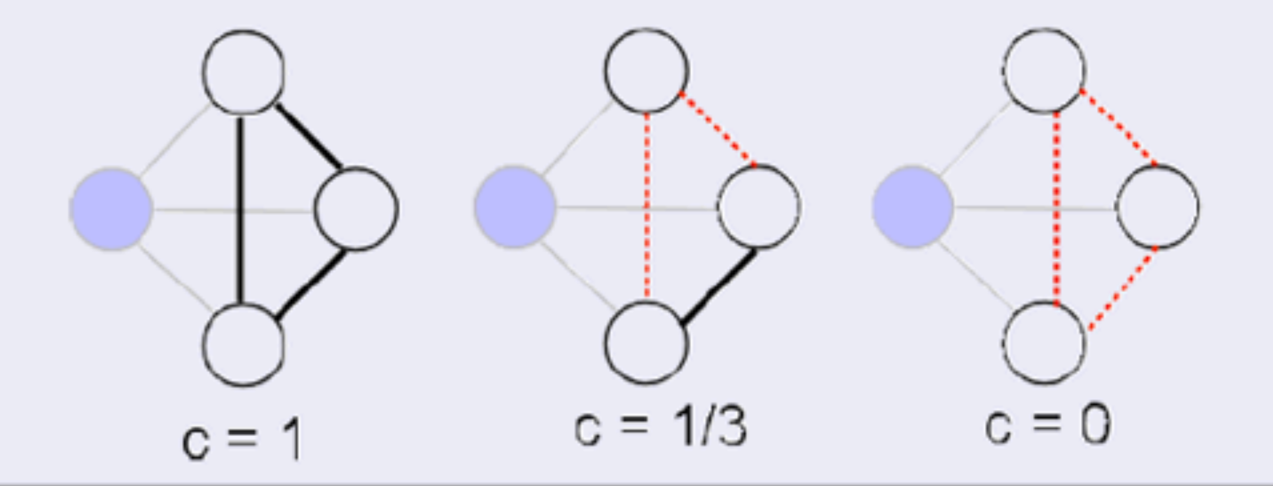
Croissance: modèle «attachement préférentiel»



Indices caractérisant les nœuds ou arêtes d'un RS

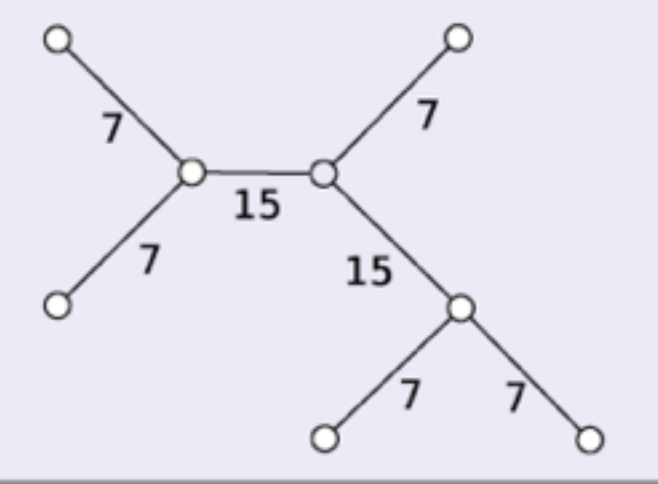
Coefficient de clustering

Lié au nombre de voisins d'un nœud qui sont eux mêmes reliés entre eux (triangles) (Watts et Strogatz, 1998)

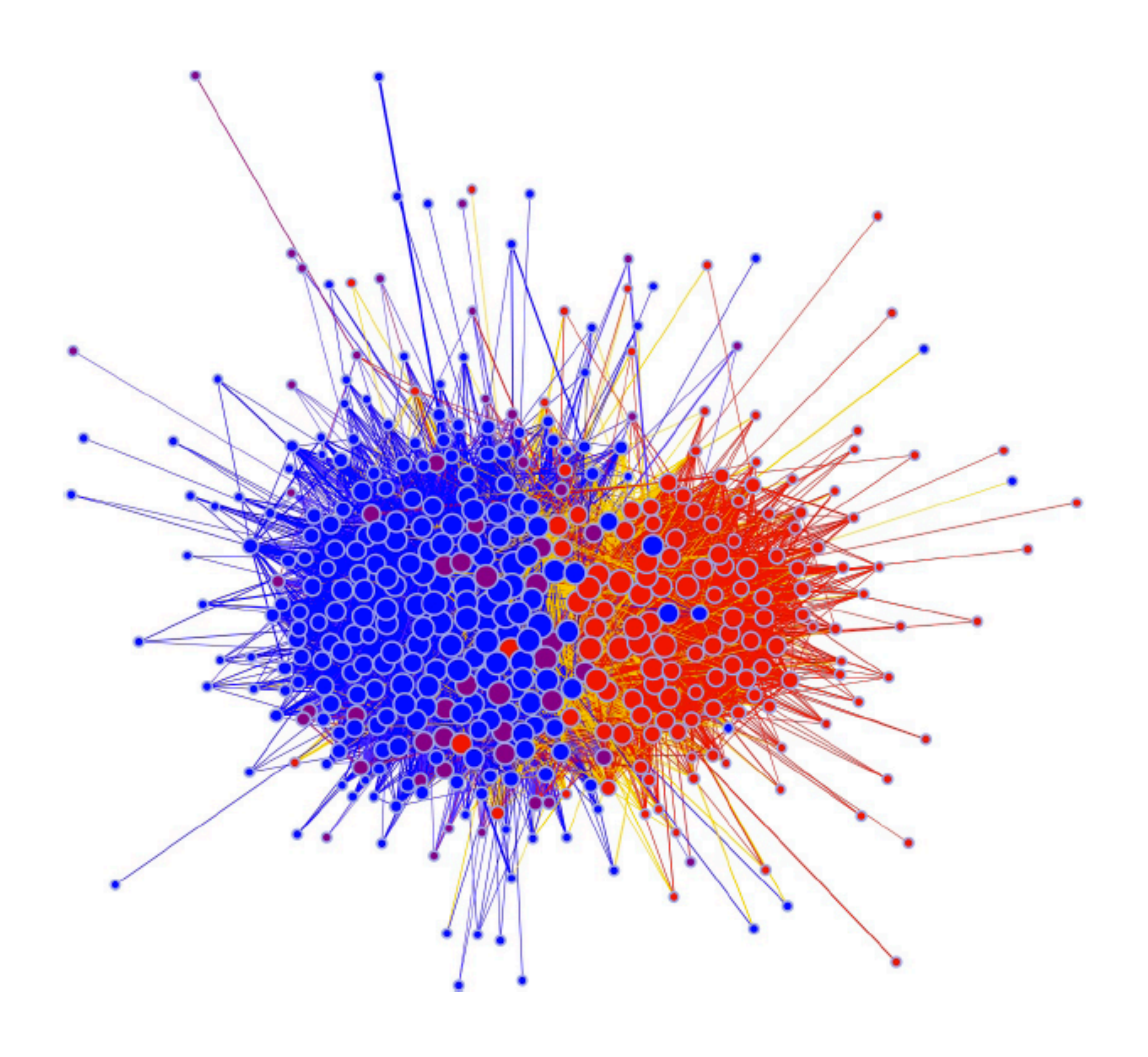


Intermédiarité

Nombre de plus courts chemins passant par une arête (Newman 2004)



Détection de communautés dans les réseaux



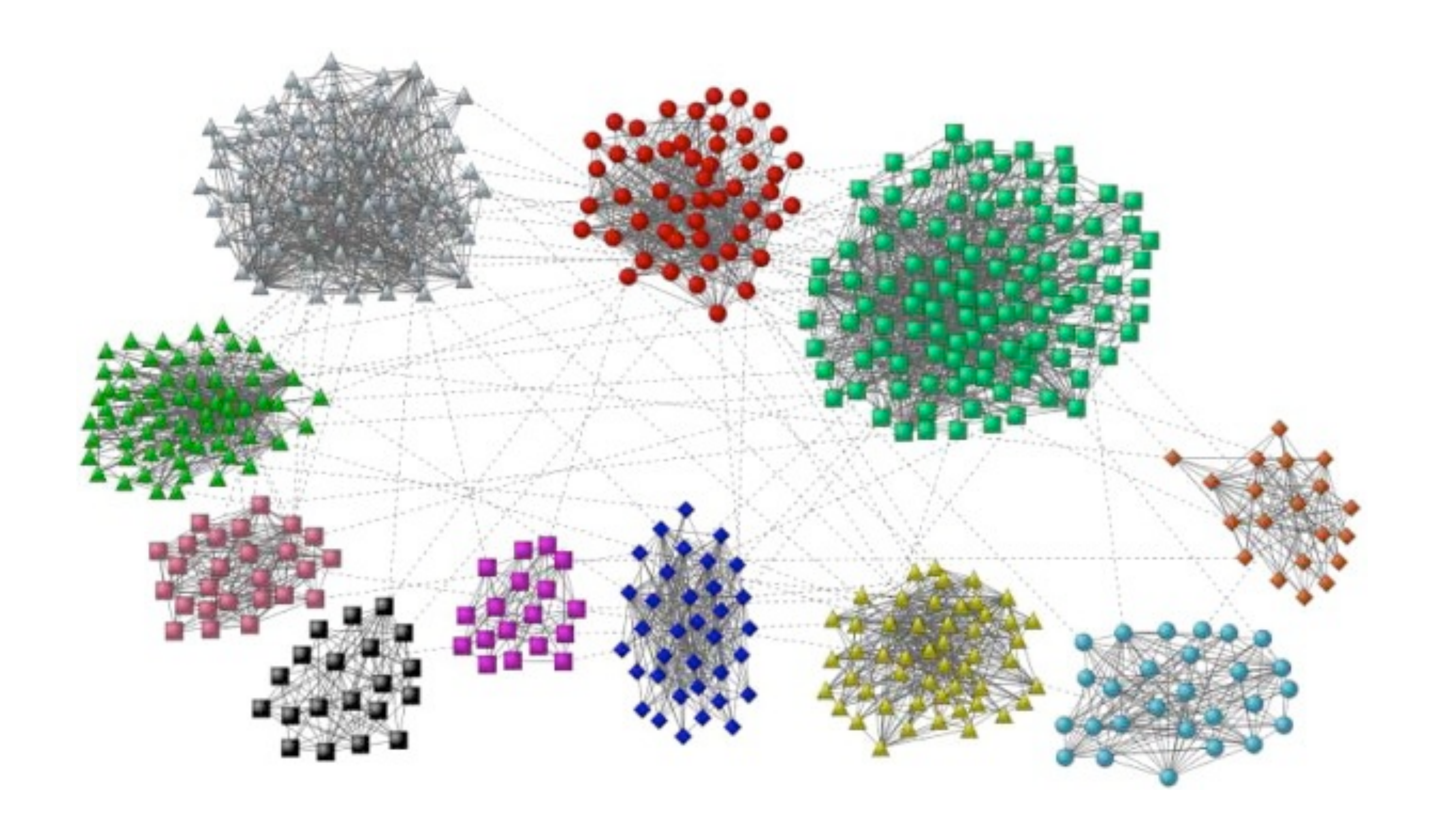
Communautés

(P. Pons, 2007)

- Recherche de communautés = partitionnement du graphe en N
- Identification = recherche d'une communauté autour d'un nœud donné



On ne considère que les liens (la «structure» du réseau)



A. Lancichinetti, S. Fortunato (2009)

Principal critère de qualité: la modularité

Newman (2004)

La modularité mesure la qualité d'un découpage du graphe en c communautés

$$Q = \sum_i (d_{ii} - (\sum_j d_{ij})^2)$$

D matrice $c \times c$, dont les éléments d_{ij} donnent la proportion de liens reliant des nœuds de la communauté i à la communauté j

$Q \in [-1, 1]$ mesure la densité des liens intra-communautaires vs inter-communautaires




Recherche partition optimale: NP complet

Recherche de communautés structurelles

De nombreux progrès récents

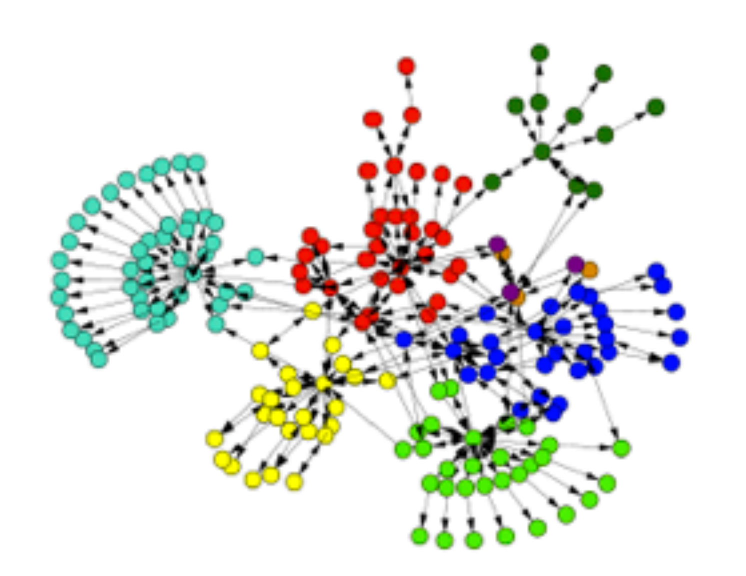
Méthodes basées sur l'*intermédiarité*

Première proposition: Newman & Girvan (2004)

- Répéter:
  1. calculer l'intermédiarité des arêtes
  2. couper l'arête la plus importante
- jusqu'à isoler tous les nœuds (méthode séparative)

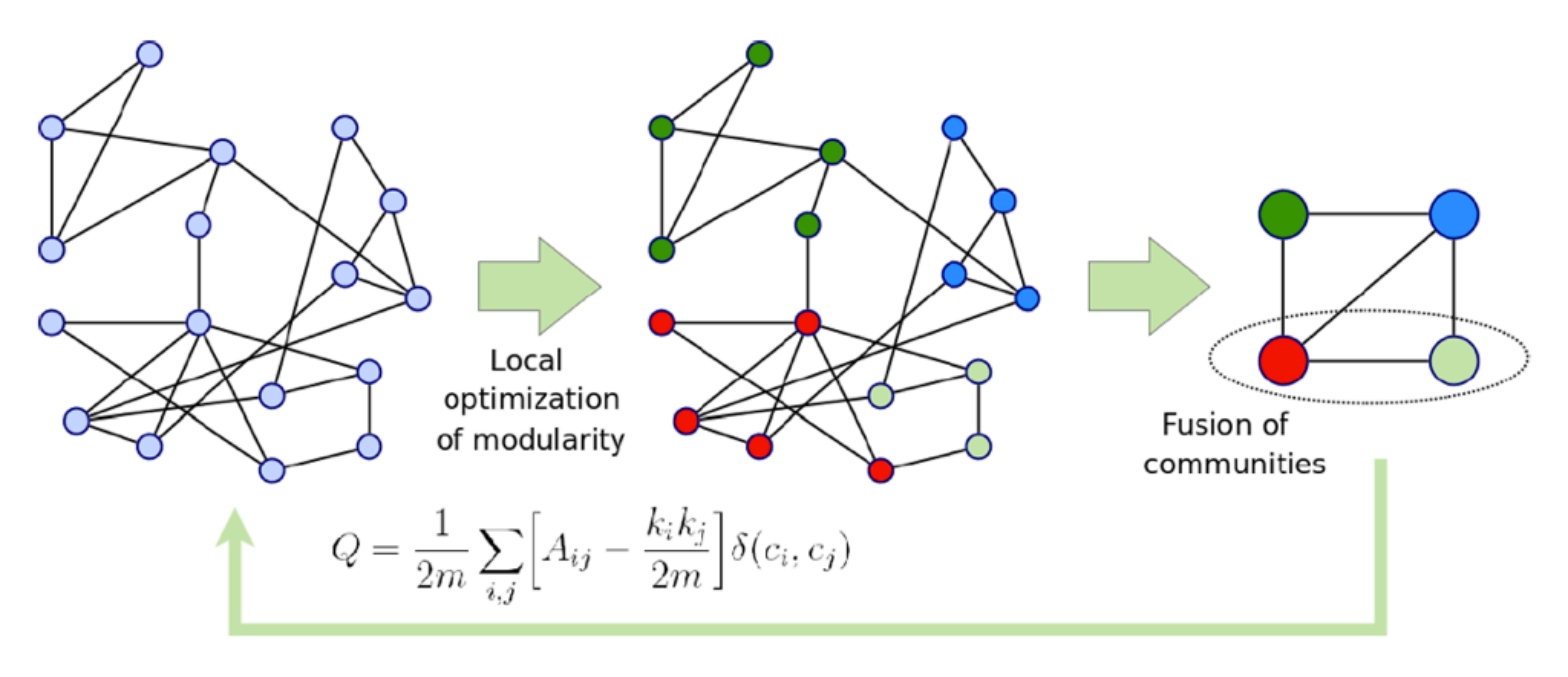
Pour un grand réseau parcimonieux de n nœuds:

Newman & Girvan	2004	$O(n^3)$
Newman	2004	$O(n^2)$
Wakita & Tsurumi	2007	$O(n \log^2 n)$
Blondel et al. (Louvain)	2008	quasi-linéaire



⇝ moins de 5 minutes pour 1 million de nœuds, ou 40 minutes pour 23 millions

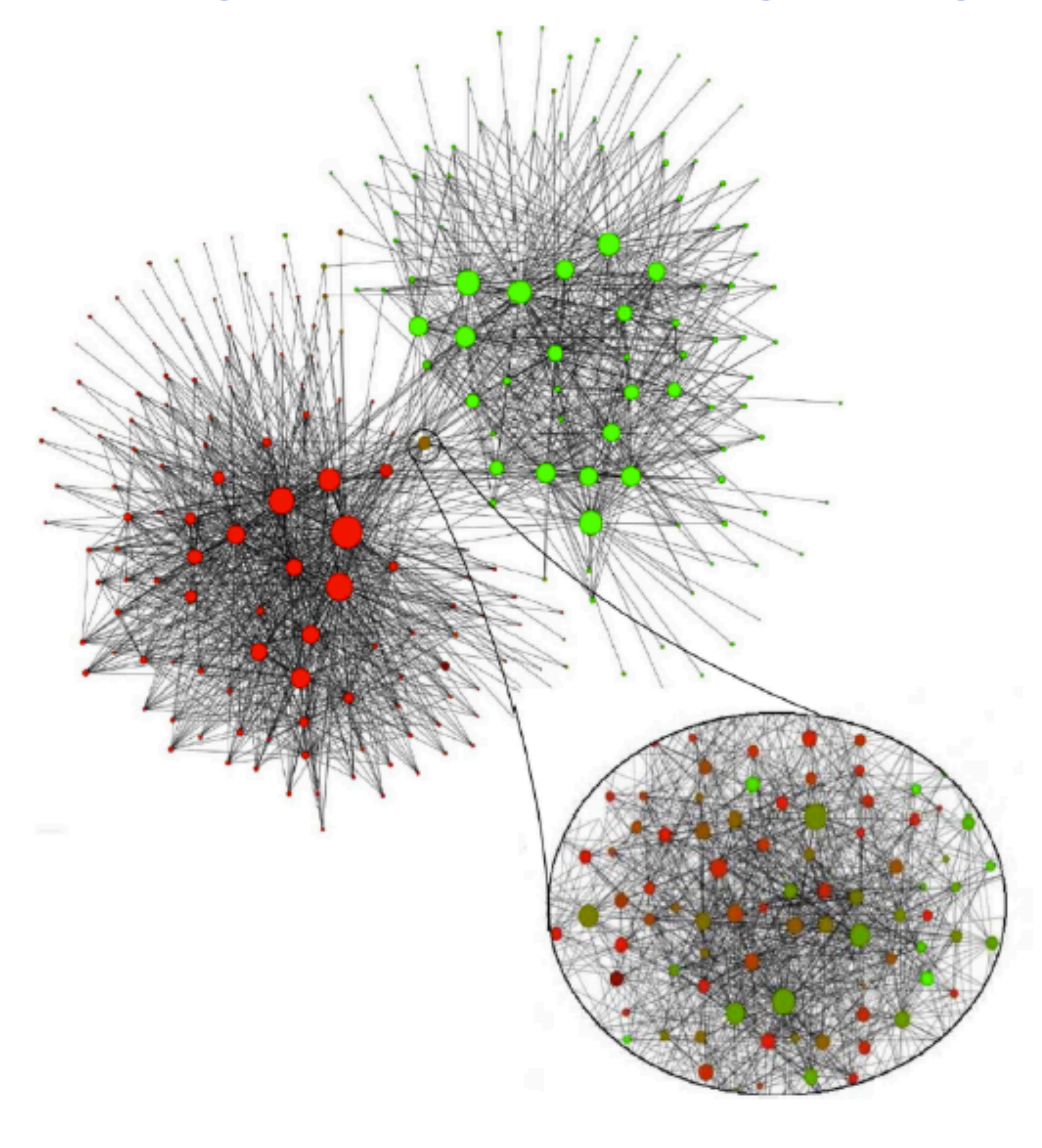
Recherche de communautés: méthode de Louvain



Optimisation locale, on tente d’associer un nœud à la même communauté que l’un de ses voisins.

Blondel et al., Fast unfolding of communites in large networks, 2008

Exemple 2: réseau téléphonique en Belgique



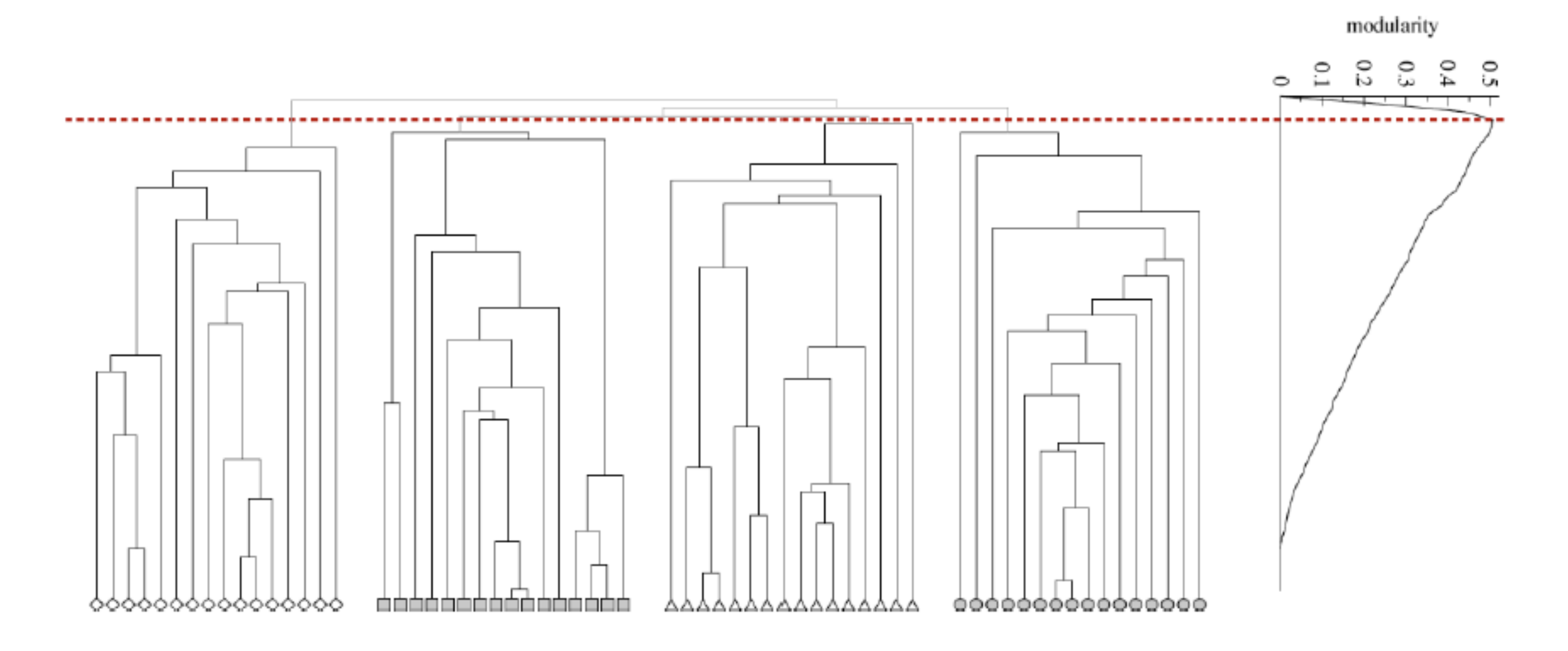
2 millions de clients

Ici chaque cercle est une communauté avec > 100 clients

Les couleurs représentent le langage parlé dans la communauté (français en rouge, flamand en vert)

D'après Blondel, et al.

Hiérarchie de communautés et modularité



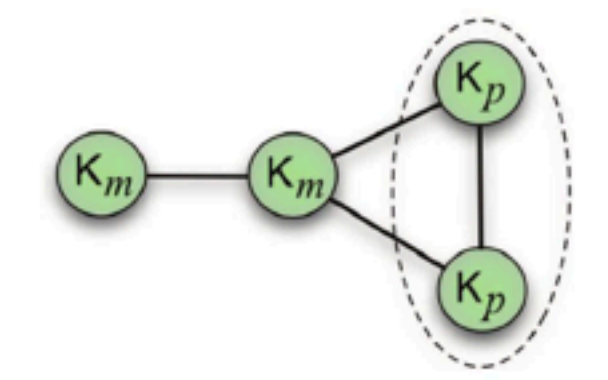
D'après Newman & Girvan, 2004

Limitations des approches basées sur la modularité

- Stabilité: de nombreuses partitions donnent des valeur proches de la modularité

- Pas de critère fiable sur l'existence ou non de structure communautaire

- Résolution limite (détection des petites communautés difficile : $m_c < \sqrt{m}$) [Fortunato 2007]



(emprunt à [Massoud Seifi 2012])

Soit K_x une clique de taille x. Si $p << m$, les deux cliques K_p sont réunies en une seule communauté, bien qu'il n'y ait qu'un seul lien entre elles

Piste: formes fortes («cœurs de communautés»)

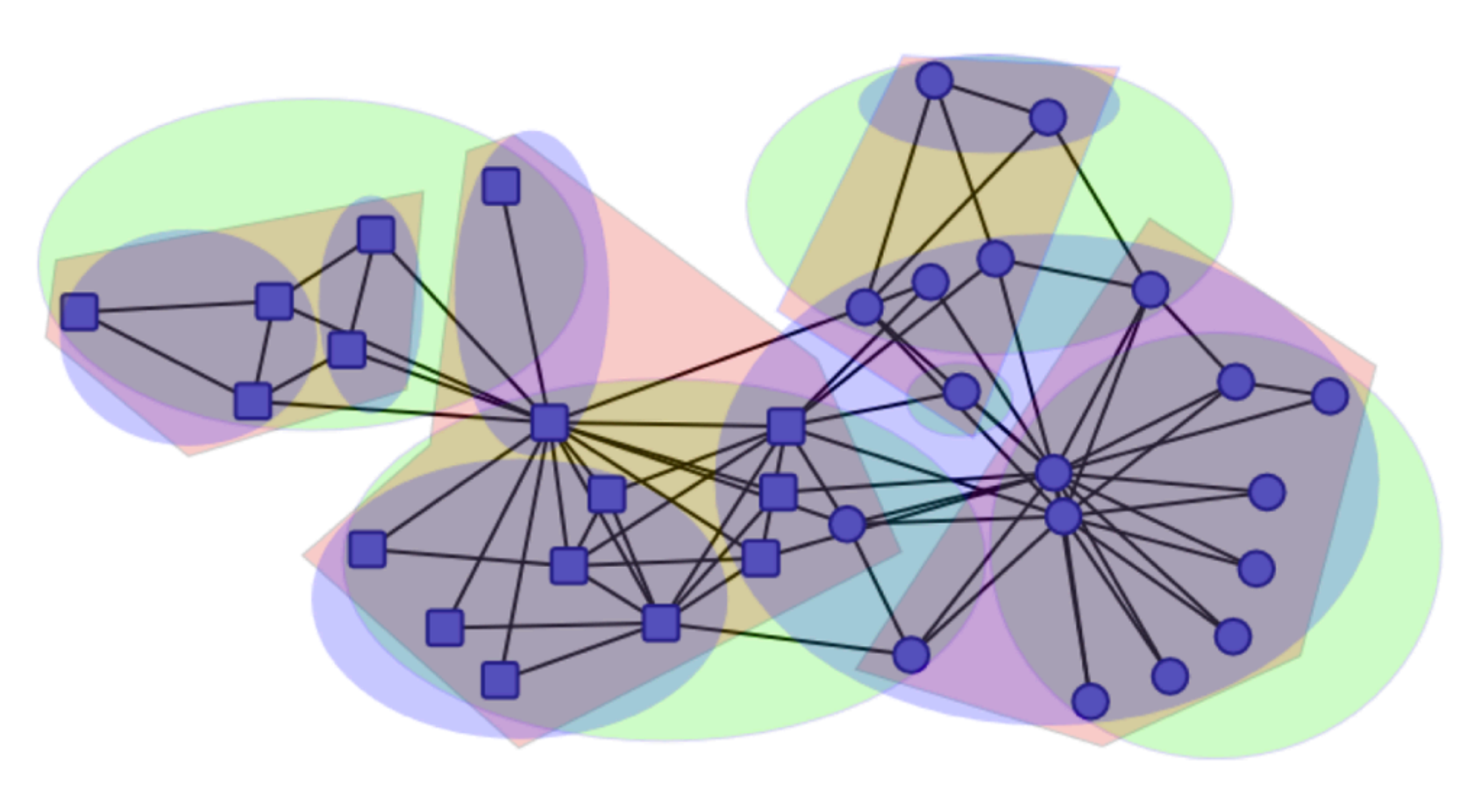
- Algorithmes d’optimisation non déterministes (ex: Louvain)
- Combiner différentes partitions

➡ Lancichinetti & Fortunato, Consensus clustering in complex networks, *Nature* (3/2012)

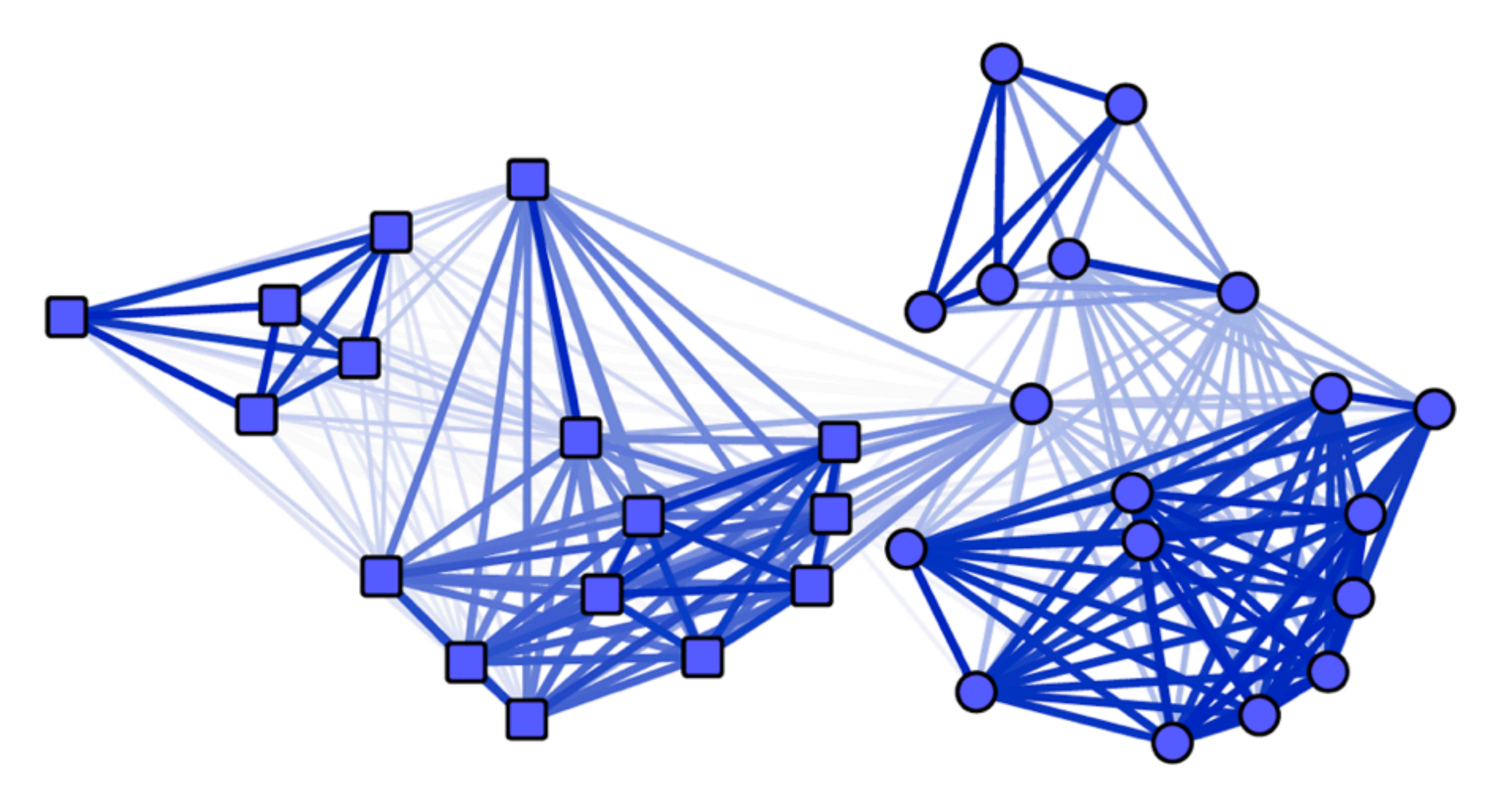
➡ Seifi, Cœurs stables de communautés dans les graphes de terrain, Thèse LIP6, 2012

(figures empruntées à [Massoud Seifi 2012])

3 partitions différentes



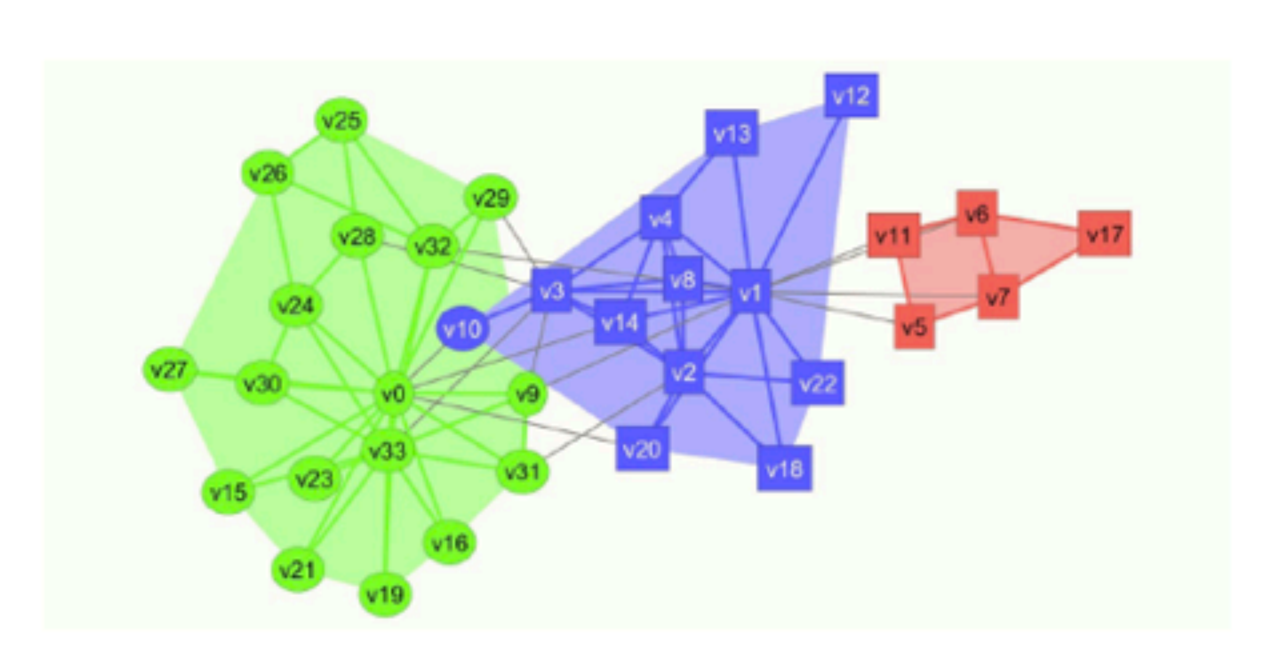
Graphe pondéré: co-occurrences dans les communautés



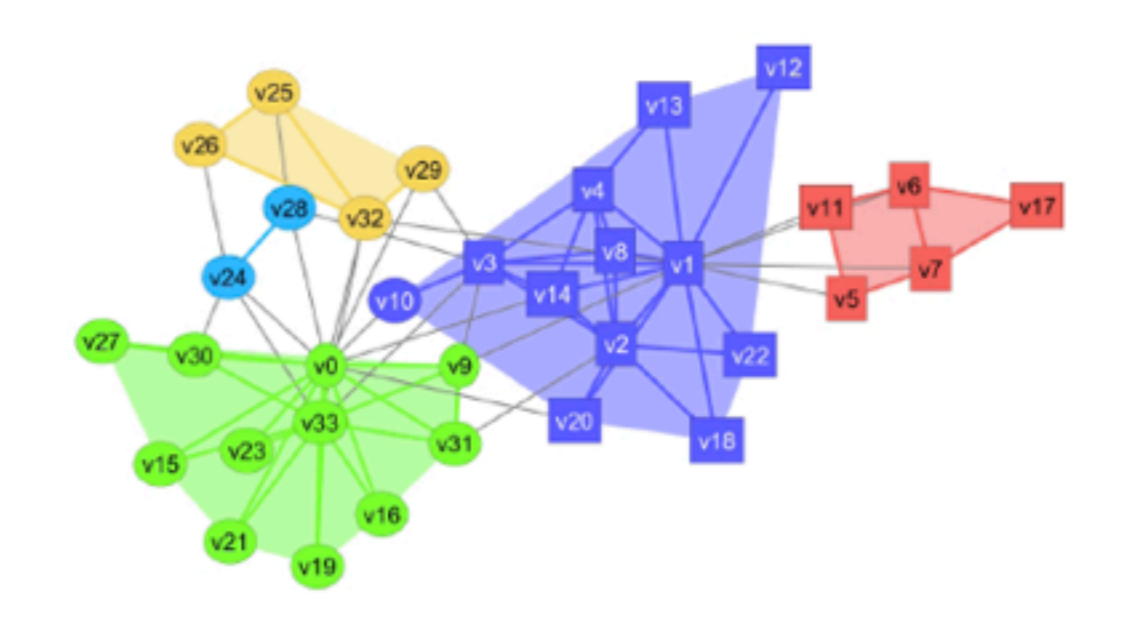
$$G' = (V, E', P^{\mathcal{N}}), \mathcal{N} = 1000$$

(figures empruntées à [Massoud Seifi 2012])

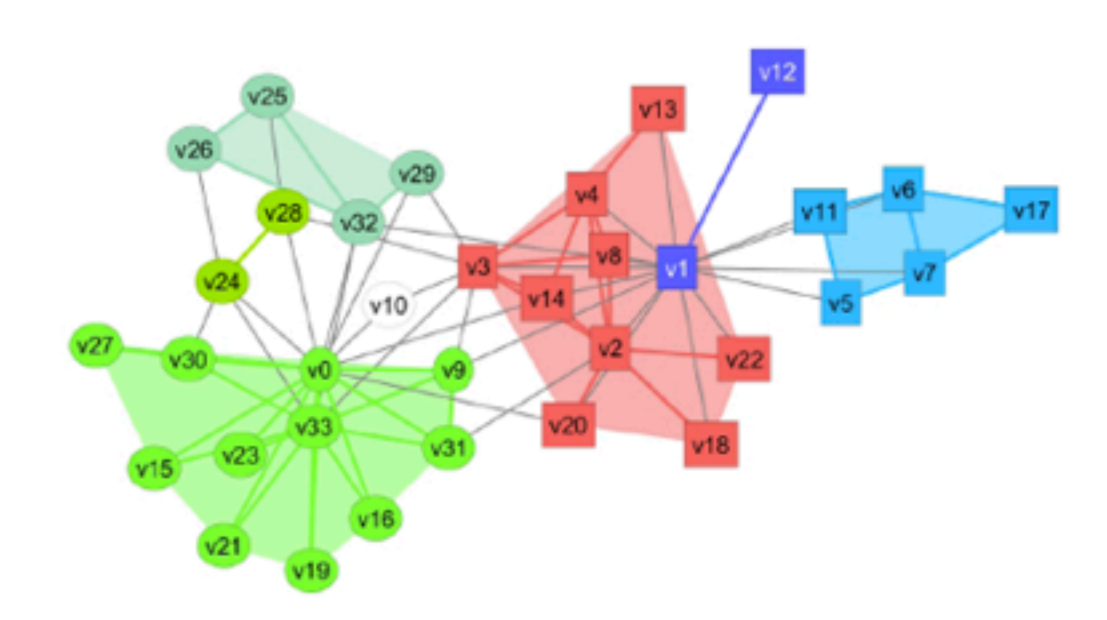
«Cœurs» : composantes connexes du graphe précédent, avec seuil \alpha



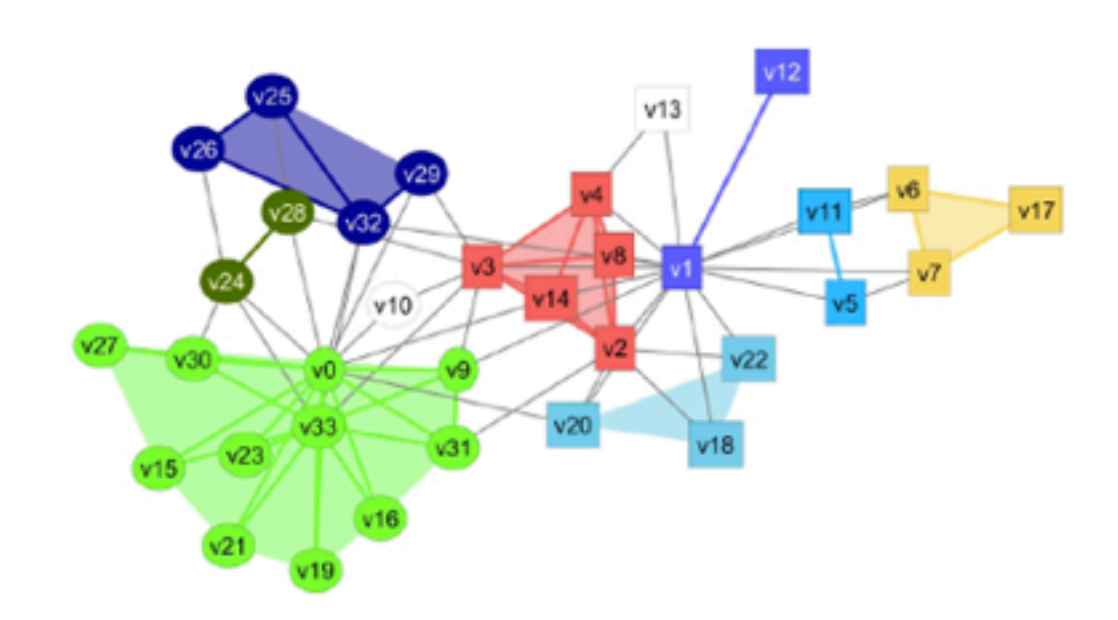
(a) $\alpha = 0.32$



(b) $\alpha = 0.62$



(c) $\alpha = 0.96$

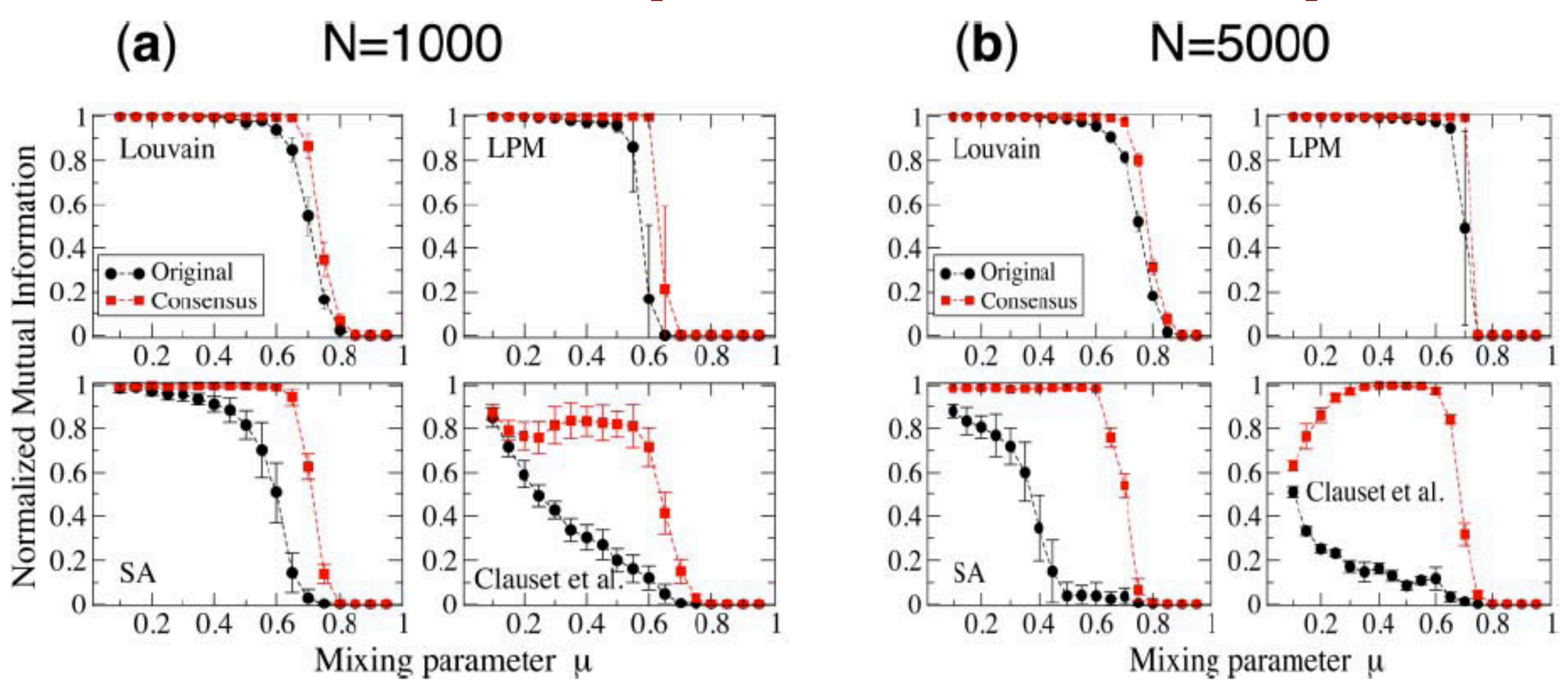


(d) $\alpha = 1.00$

Cœurs identifiés en utilisant quatre seuils différents.

«Cœurs» : résultats plus précis sur graphes synthétiques (benchmark LFR)

[Lancichinetti & Fortunato 2012]



- Stabilité: possibilité de suivi temporel
- Comportement en fonction du seuil alpha: permettrait de décider si structure communautaire (?)

Approches de type «consensus»

- Stabilité: possibilité de suivi temporel
- Comportement en fonction du seuil alpha: permettrait de décider si structure communautaire (?)
  (étude des phénomènes de transition de phase)

... mais 100 à 1000 fois plus lent
... pas de prise en compte des attributs

Prise en compte des attributs des nœuds

Prise en compte des attributs des nœuds

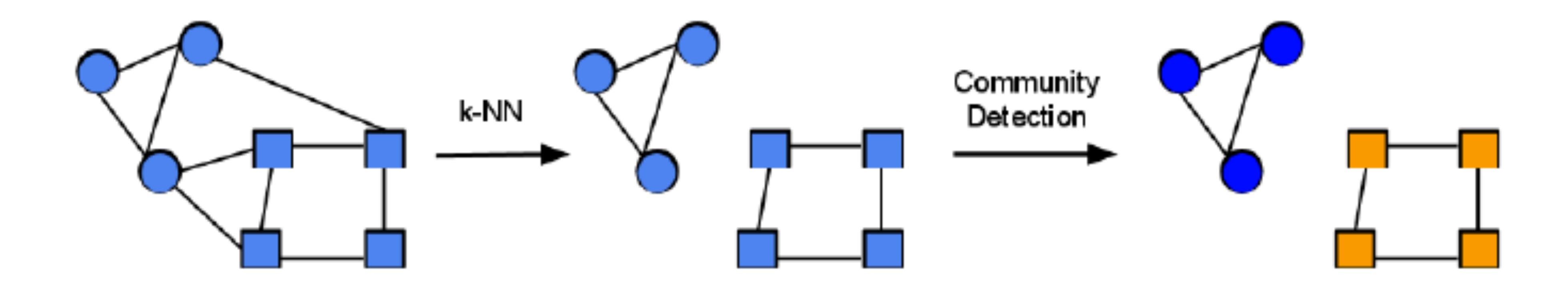
Une idée simple: (méthode «SAC2», Dang & Viennet 2012)

- on se donne une mesure de similarité *simA* entre nœuds (basée sur les attributs portés par ces nœuds)
- on définit une nouvelle mesure de similarité basée sur lien (pondéré) et *simA*

Similarity measure of 2 nodes i and j:

$$S(i,j) = \alpha \cdot G_{i,j} + (1-\alpha) \cdot simA(i,j)$$

- Création d'un graphe de voisinage (kNN), et partitionnement de ce graphe.



Phase 1

Similarity measure of 2 nodes i and j:

$$S(i,j) = \alpha \cdot G_{i,j} + (1-\alpha) \cdot simA(i,j)$$

Phase 2

- k-nearest neighbor graph (k-NN) $G_k = (V, E_k)$: Node i and j are connected if $i \in kNN(j)$ or $j \in kNN(i)$, $kNN(i)$ denotes k nearest neighbor of i
- Exists methods that approximate k-nearest neighbor graph with lower complexity, i.e., $O(n^{1.14})$ (*Dong et. al (2011)*)

Phase 3

Louvain algorithm is used to find the communities

Comparison of Methods

Extract the communities from the above datasets, using different methods :

Attribute-only

- Attribute-only clustering: K-means method is used to group nodes based on the similarity in attributes (link information is ignored)

Structure-only

- Louvain algorithm on unweighted graph *D. Blondel et al.(2008)*

Mixed

- Random walks: Method proposed by *Steinhaeuser et Chawla (2009)*, based on random walks and hierarchical clustering
- Fast greedy: Method proposed by *Clauset et al. (2004)* based on the greedy optimization of modularity. The graph is weighted by node attribute similarities
- Our proposed algorithms SAC1 and SAC2

- Communautés de tailles plus équilibrées

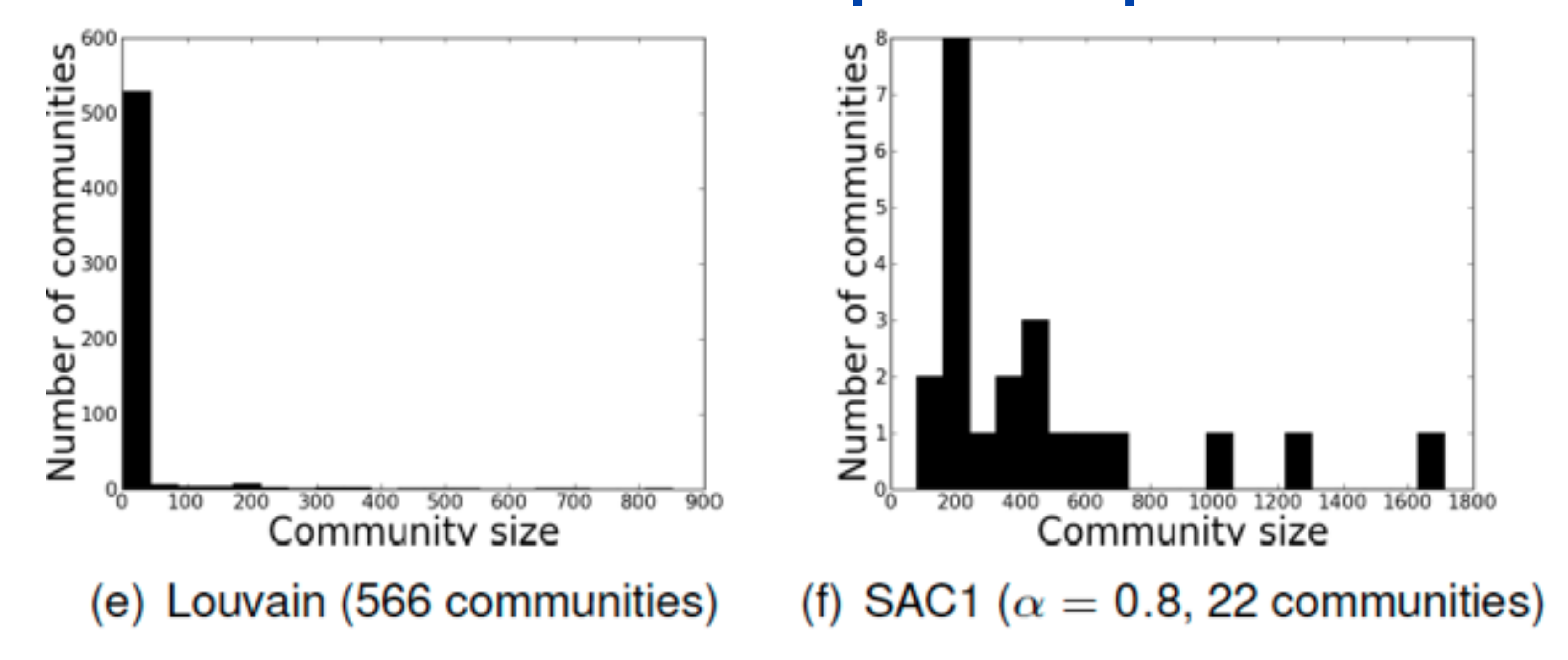
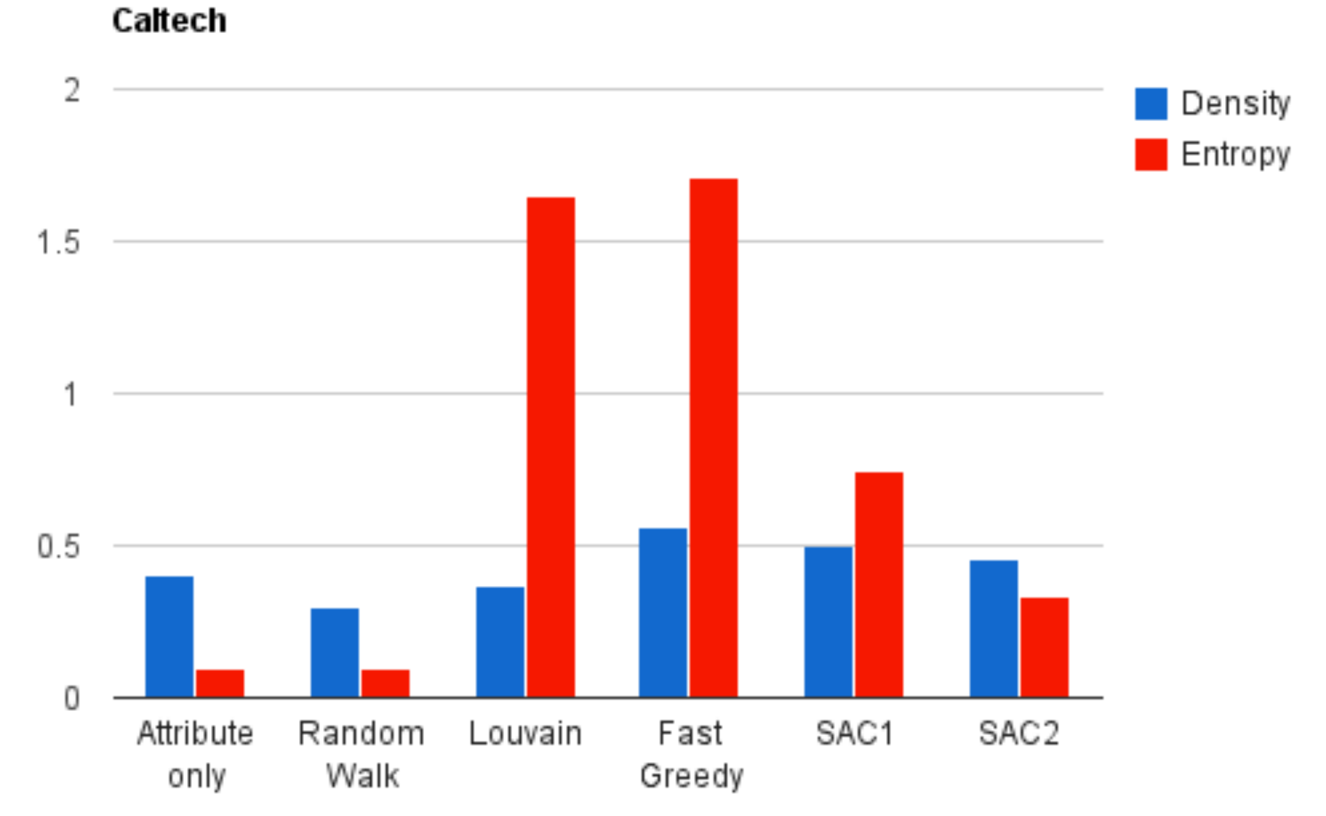


Figure: DBLP community size distribution

- Communautés homogènes et «cohésives»



Une application: blogs



Skyrock | Articles | Connecte-toi | Inscris-toi

Accueil | Blogs | Profils | Chat | plus ▼

Accueil | Blogs Stars | Top 100 | Top Kiff | Les + vus | Top 7 jours | Thèmes | Officiels | Sources

Beautiful Friend
Un source Pattinson de toute beauté. En mode Bel Ami pour l'occasion... Voir »

Connecte-toi !
Pseudo :
Mot de passe :
Rester connecté(e)
J'ai oublié mon mot de passe
Se connecter

Découvre Skyrock
Crée ton blog

Blogs Stars — Deviens Blogstar | Suite »

Tour-De-France-a-Pied
C'est le pari fou d'un homme qui marche 1500 km à pied contre la pauvreté !

IndispensableDepotoir
Elle apaise ses maux avec des mots. Demo...

JuliettaM
Suis l'actu de la plus emblématique des candidates de SS5 !

spa-de-redon35600
Le témoignage d'une bénévole à la SPA...

xx-the-world-people-xx
Une sélection des meilleures sorties en salle...

lluislacondeguy
Quand une fan de Lluis Lacondeguy partage sa passion des dirteurs et des freeriders...

Recherche blog
OK

Skyrock FM — Le blog de la FM »

Officiels (publicité) — Suite »

Les Kaïra
Les Kaïra enfin le film !

Le Dernier Rite
Le Dernier Rite dispo en DVD et BLURAY le 1er Juin

THE DARK KNIGHT RISES
Le 25 juillet au cinéma

Projet ANR ExDEUSS / CEDRES (2009-2012)

Blog | Profil

100% lluis lacondeguy- fan source français et bientôt Español

SOURCE



lluislacondeguy

Description :

voici mon blog sur mon rider dirt slopestyle mtb (et bmx) préférée , il est espagnol , talentueux et cool ^^ . je l ai découvert il y a 5 ans est maintenant je suis une vrai fans donc voila je vous fait découvrir ma passion =) j'éspére que vous aimerait les news, les images et que vous me laisseré un pti commentaire de temps en temps ^^ , bonne visite !

aquí está mi blog en mi jinete tierra mtb slopestyle (y bmx) preferida, que es el español, con talento y ^ ^ fresco. He descubierto hace unos 5 años

bienvenue ! bienvenida!

bienvenue ! bienvenida!

hello tout le monde !!!

Les raisons pour lesquels j'ai crée ce blog sont diverses et varié , la première est que je suis totalement admiratif des dirteur et freerider en général , je trouve que toute les figures qu'il font sont totalement bluffante . J aime particulierment Lluis car il me semble etre un gars super simple et extra doué et j'adore sa ! donc voila de

Infos

Création : **28/07/2009 à 14:01**
Mise à jour : **Hier à 12:08**
76 articles
80 commentaires
199 amis
16 favoris
405 kiffs

Son morceau préféré



RIZZLE KICKS
Traveller's Chant -Rizzle Kicks (Stereo Typical)



Ajouter à mon blog

Skyrock Music

Tags

... 2012 andreu lacondeguy artbmx autriche bienve aguado alba facebook page fans FISE 2011 lacondeguy ll lluis lluis lacondeguy master of dirt new world disorder patrick quimez photoshoot sam reynolds sebas romero photography tj ellis

» Suite

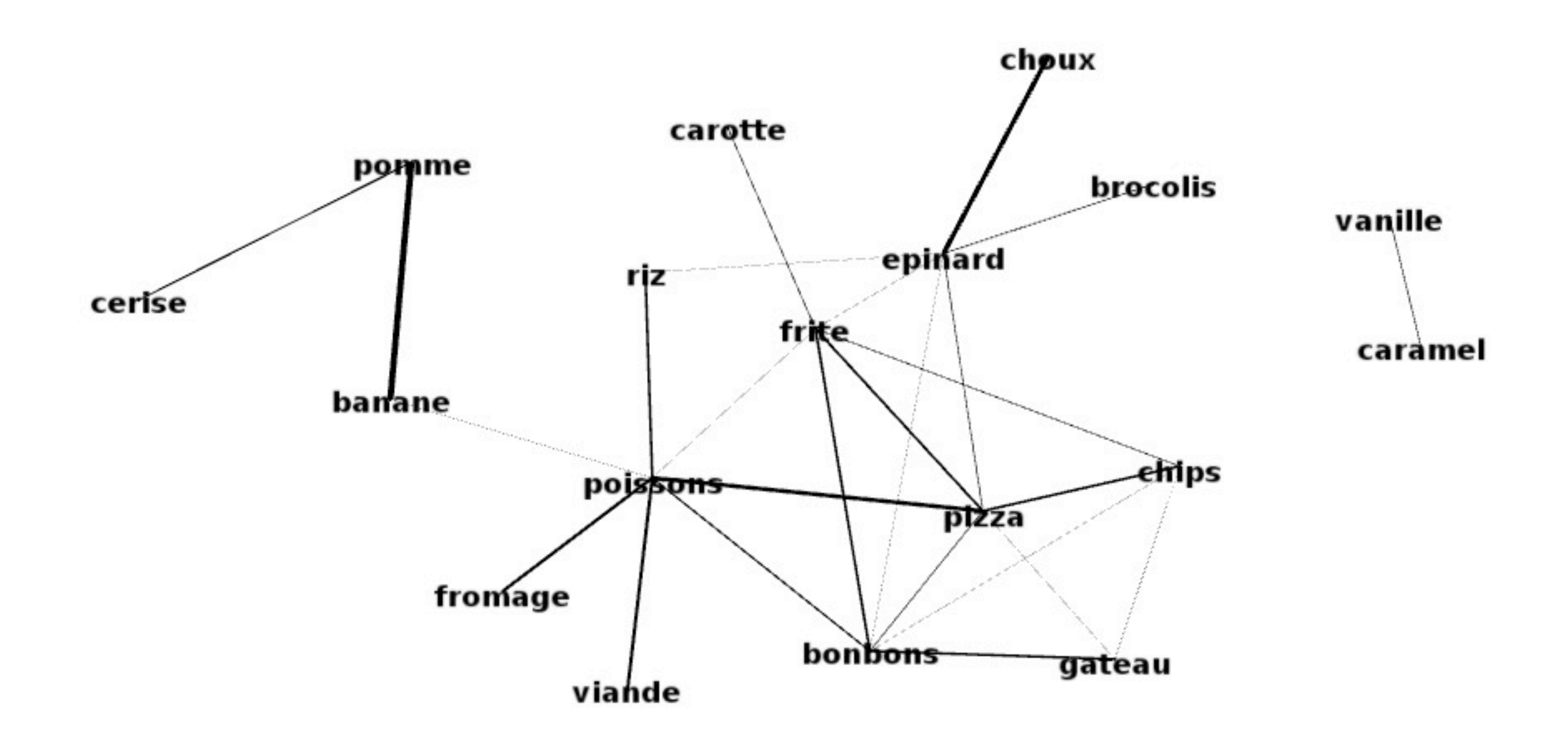
A partir de cela, on peut créer un réseau bipartite

utilisateurs x étiquettes (*tags*)

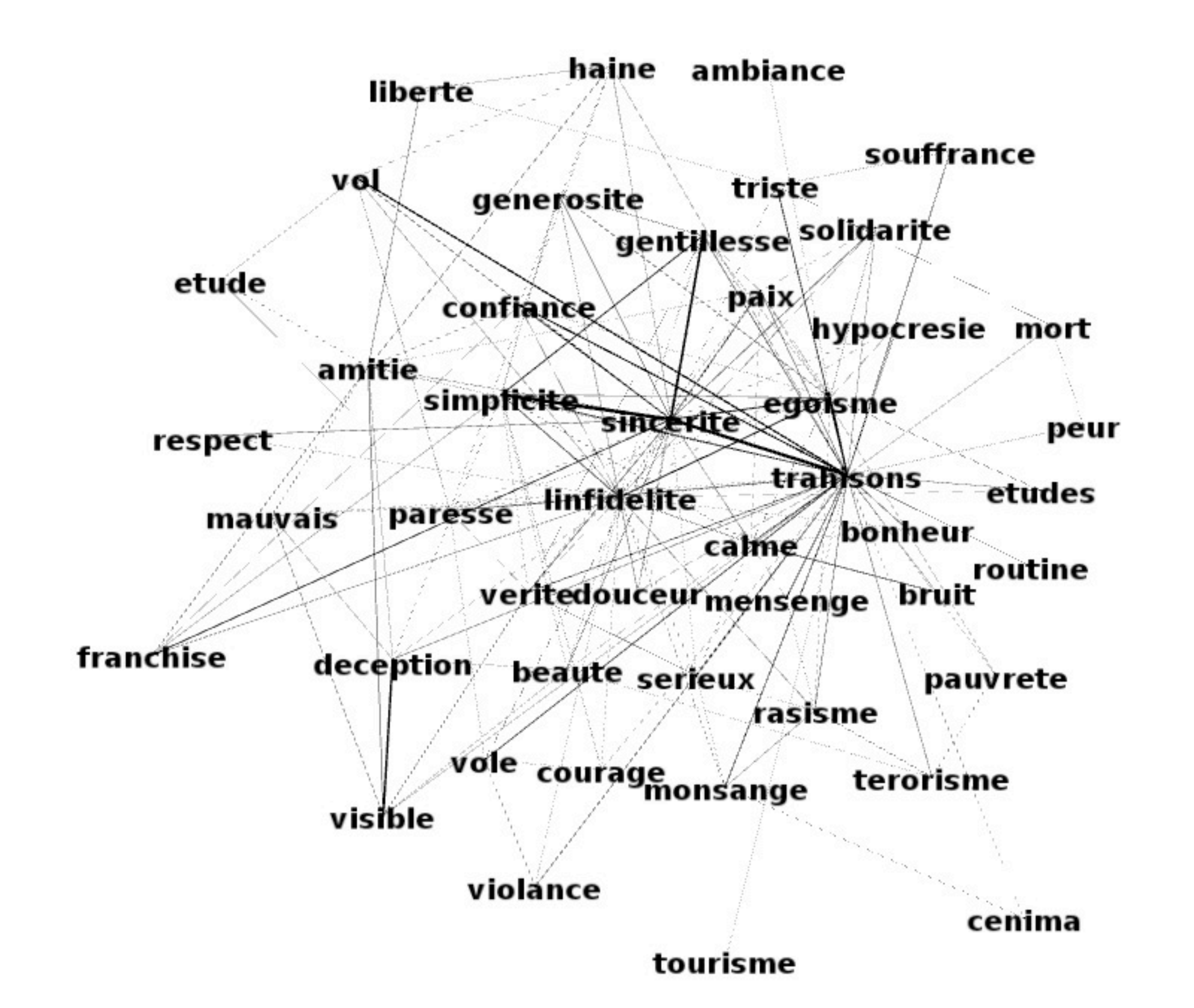
Et le projeter côté étiquettes, pour découvrir des thématiques...

- Deux étiquettes sont connectées si
  elles apparaissent ensembles sur un utilisateur

- Similarité entre étiquettes



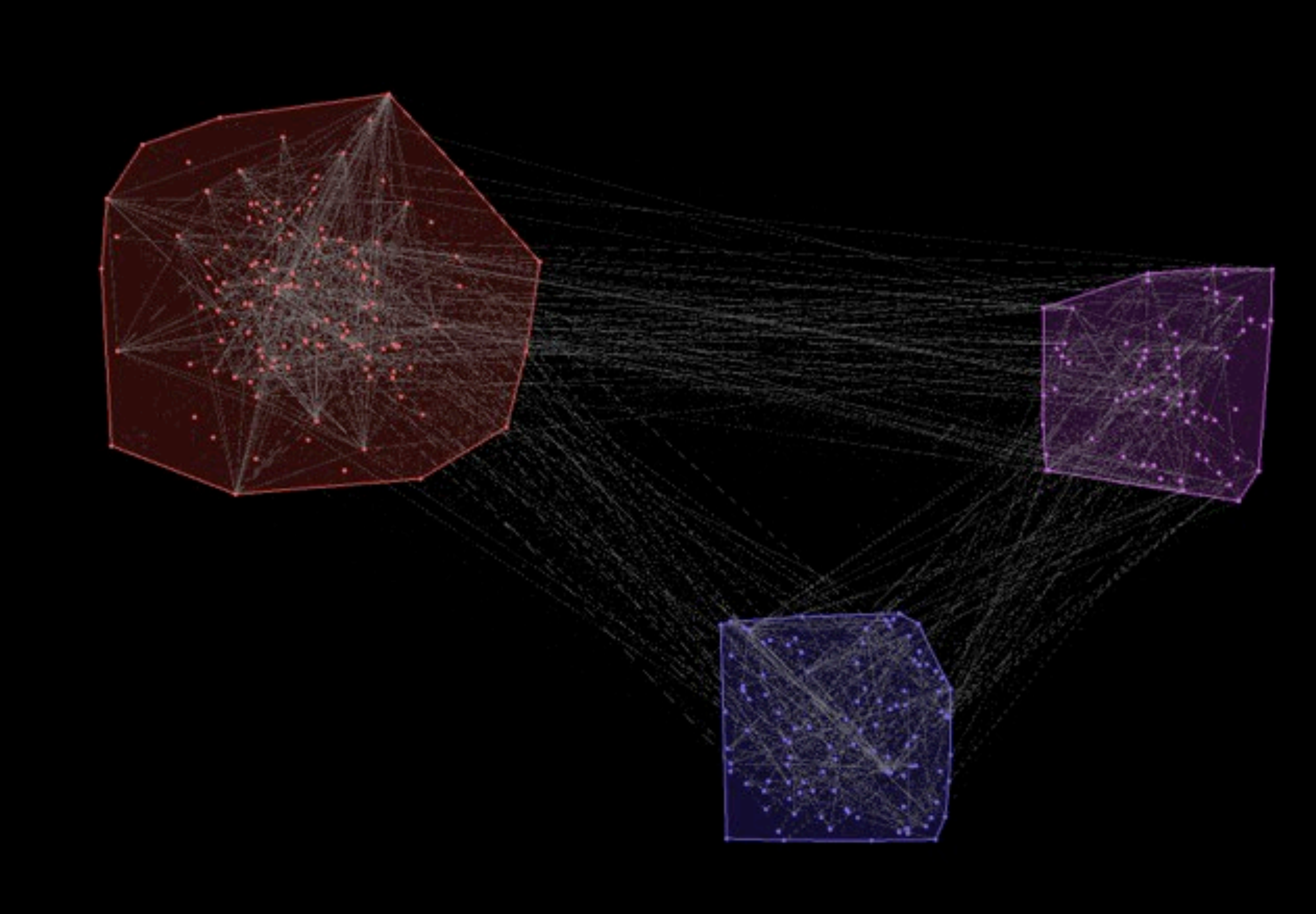
[Dang & Viennet 2012]

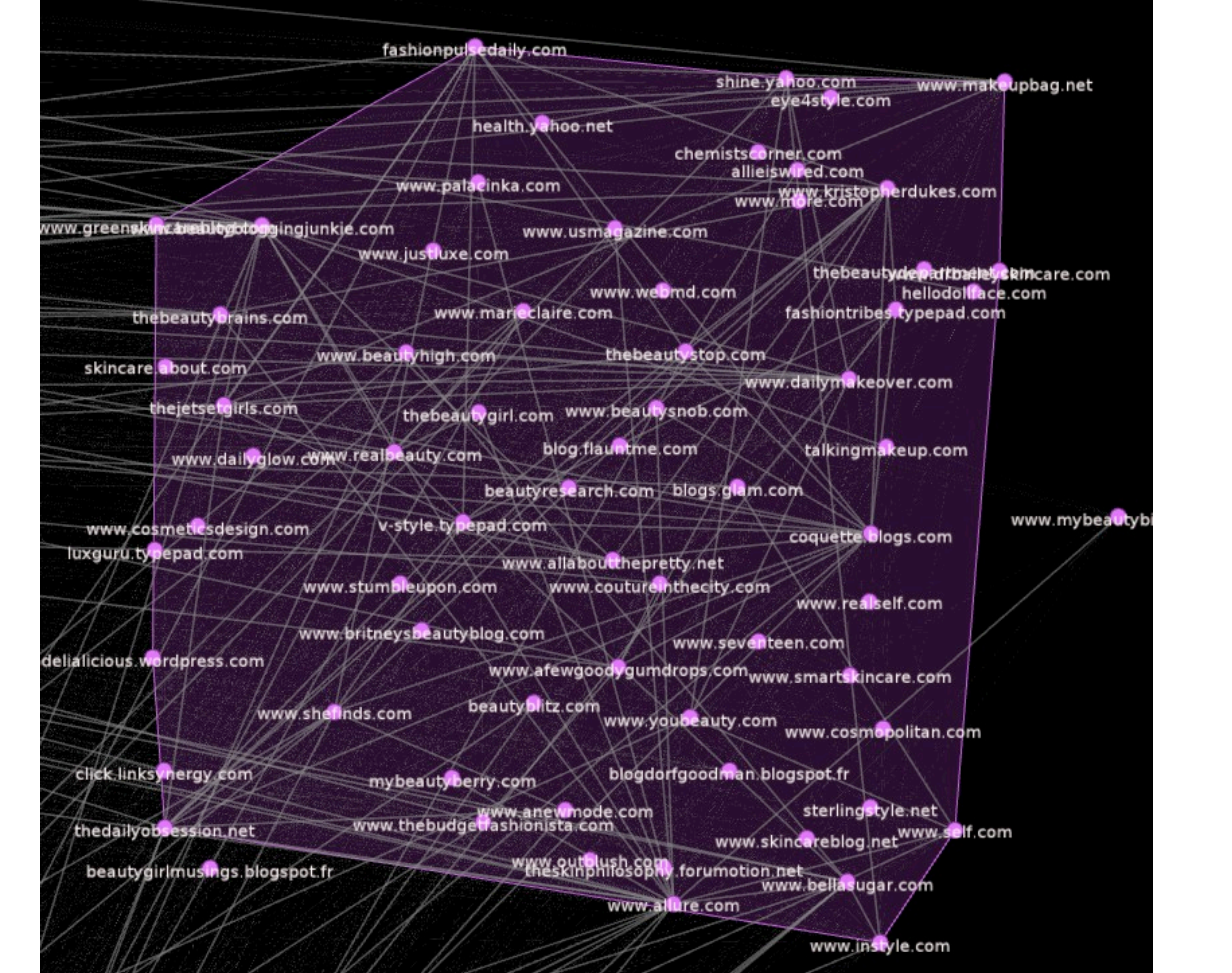

haine
ambiance
liberte
souffrance
vol
triste
generosite
solidarite
gentillesse
etude
paix
confiance
hypocresie
mort
amitie
simplicite
egoisme
sincerite
peur
respect
trahisons
etudes
mauvais
paresse
linfidelite
bonheur
calme
routine
verite
douceur
mensenge
bruit
franchise
deception
beaute
serieux
pauvrete
rasisme
vole
courage
monsange
terorisme
visible
violance
cenima
tourisme

Une autre application: réseaux de sites web (affiliés à des marques)

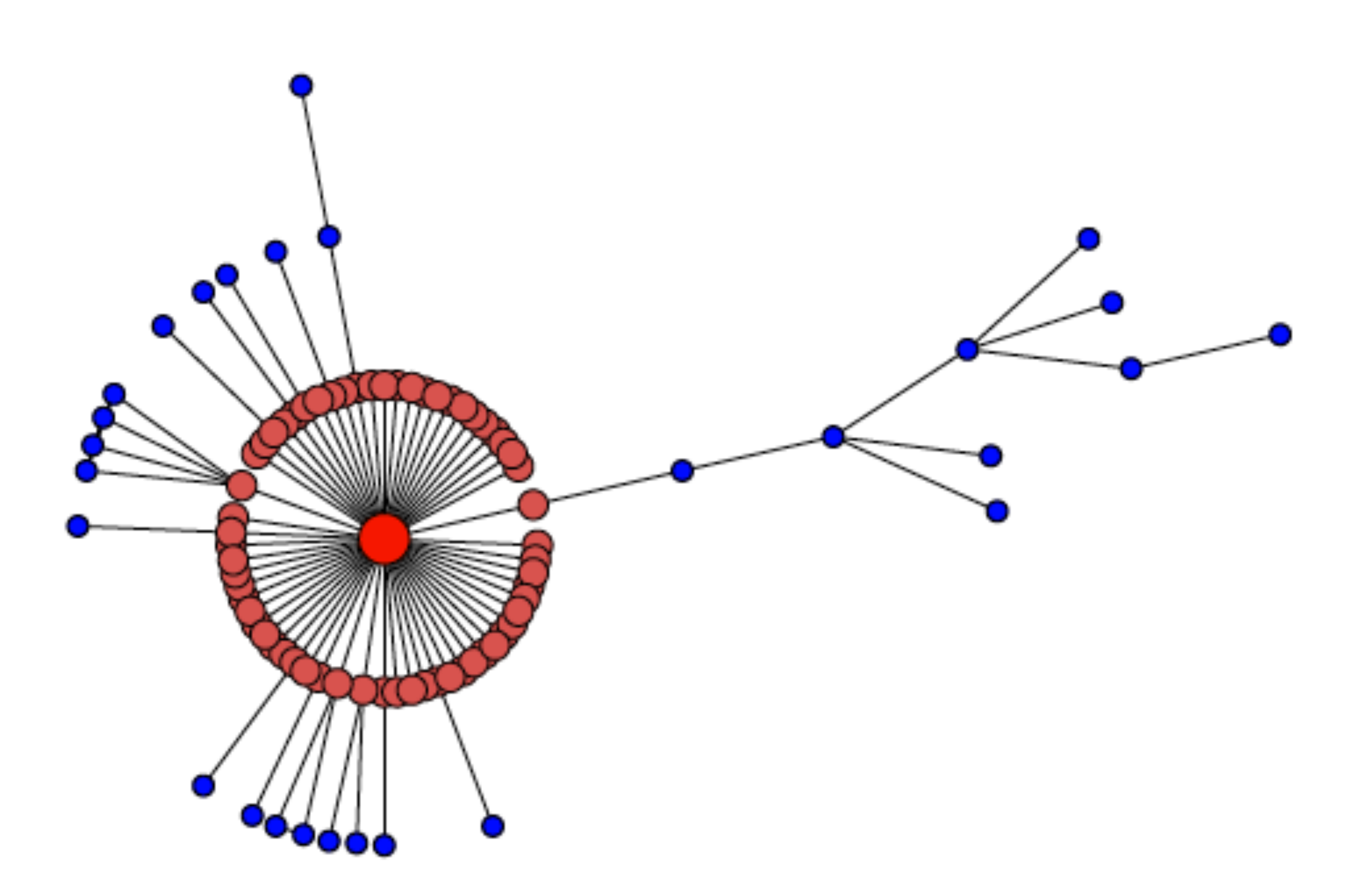
Liens (html) entre sites, et tags (sur les pages des blogs)

=> graphe avec attributs, que l'on segmente



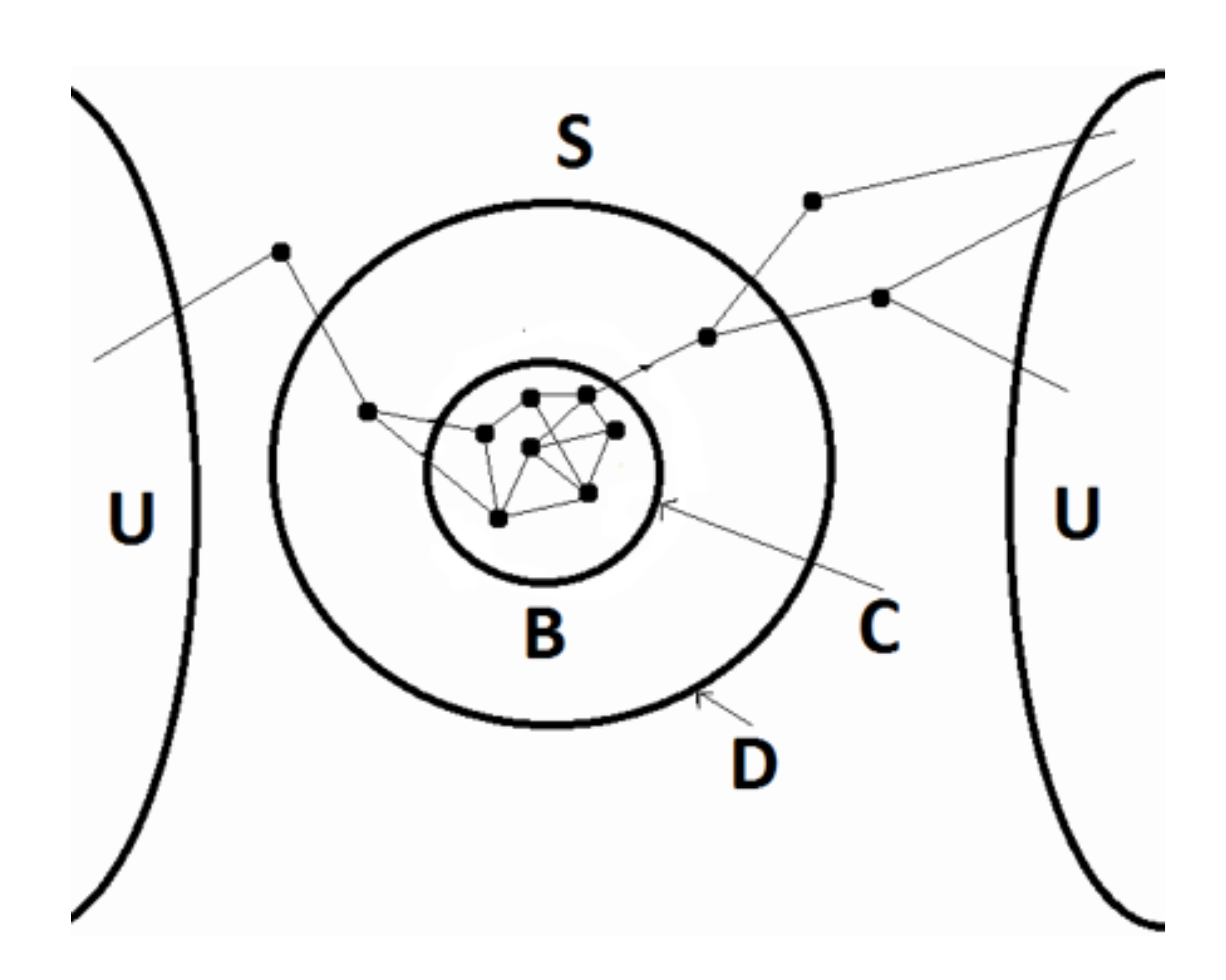

fashionpulsedaily.com
shine.yahoo.com
www.makeupbag.net
eye4style.com
health.yahoo.net
chemistscorner.com
allieiswired.com
www.kristopherdukes.com
www.more.com
www.palacinka.com
www.usmagazine.com
www.justluxe.com
www.webmd.com
hellodollface.com
thebeautybrains.com
www.marieclaire.com
fashiontribes.typepad.com
www.beautyhigh.com
thebeautystop.com
skincare.about.com
www.dailymakeover.com
thejetsetgirls.com
thebeautygirl.com
www.beautysnob.com
blog.flauntme.com
talkingmakeup.com
www.dailyglow.com
www.realbeauty.com
beautyresearch.com
blogs.glam.com
www.cosmeticsdesign.com
v-style.typepad.com
coquette.blogs.com
luxguru.typepad.com
www.allaboutthepretty.net
www.stumbleupon.com
www.coutureinthecity.com
www.realself.com
www.britneysbeautyblog.com
www.seventeen.com
delialicious.wordpress.com
www.afewgoodygumdrops.com
www.smartskincare.com
www.shefinds.com
beautyblitz.com
www.youbeauty.com
www.cosmopolitan.com
click.linksynergy.com
mybeautyberry.com
blogdorfgoodman.blogspot.fr
www.anewmode.com
sterlingstyle.net
thedailyobsession.net
www.thebudgetfashionista.com
www.self.com
www.skincareblog.net
beautygirlmusings.blogspot.fr
www.outblush.com
theskinphilosophy.forumotion.net
www.bellasugar.com
www.allure.com
www.instyle.com

Communautés locales

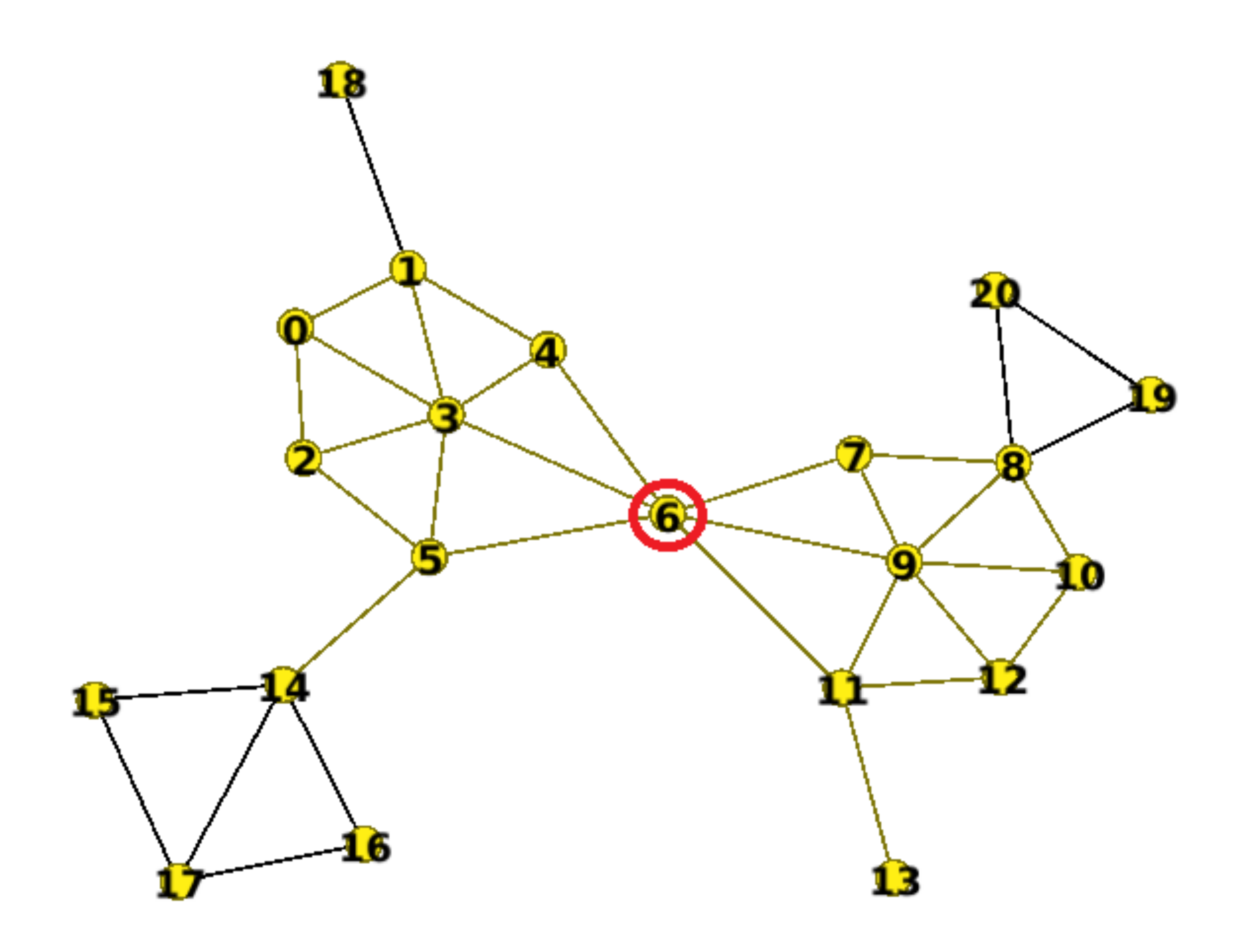


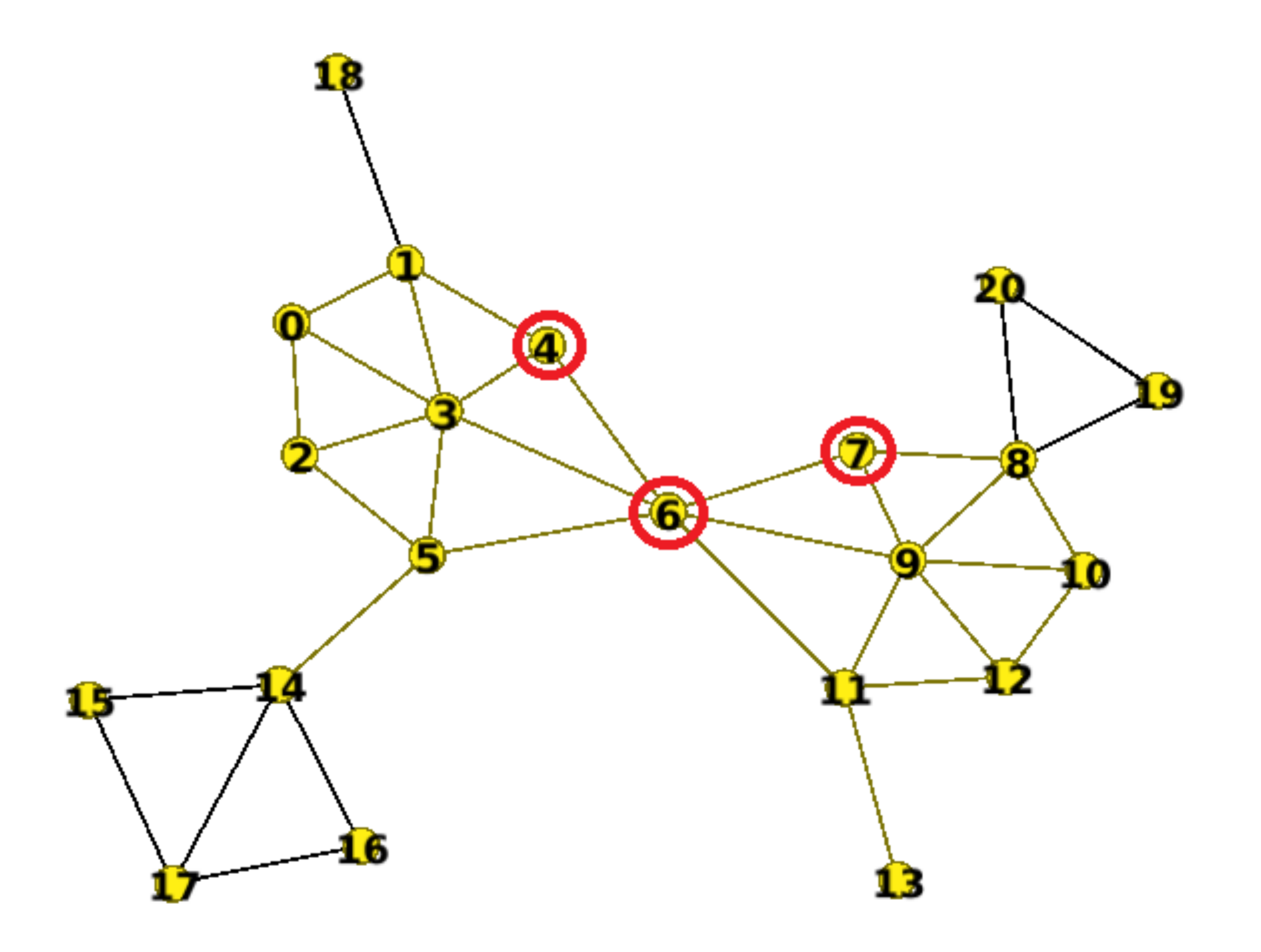
Algorithme de détection de la communauté locale d'un nœud
(Ngonmang, Tchuente, Viennet 2012)

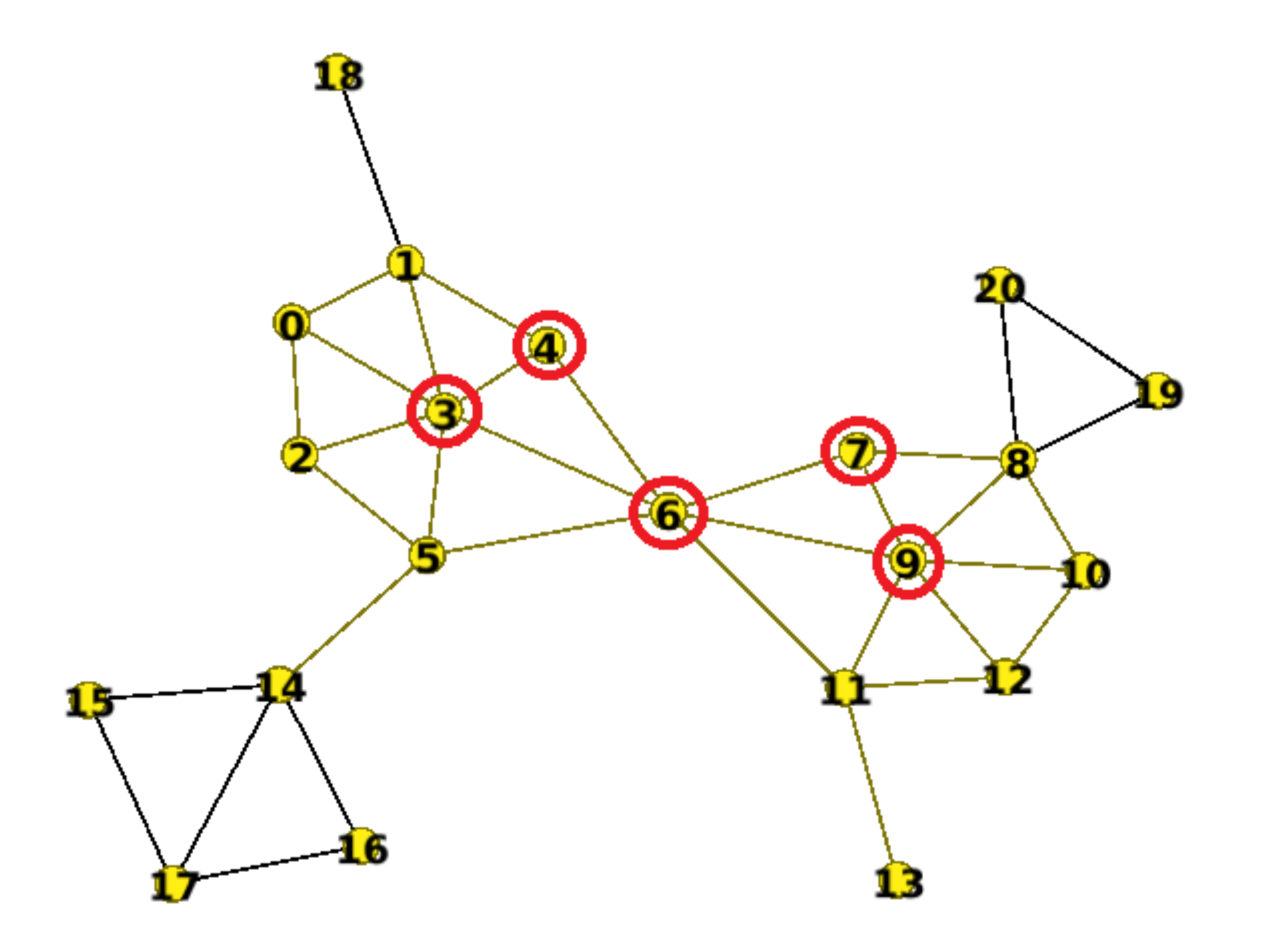
Amélioration des algorithmes de Clauset (2005), Luo (2006) et Chen (2009)

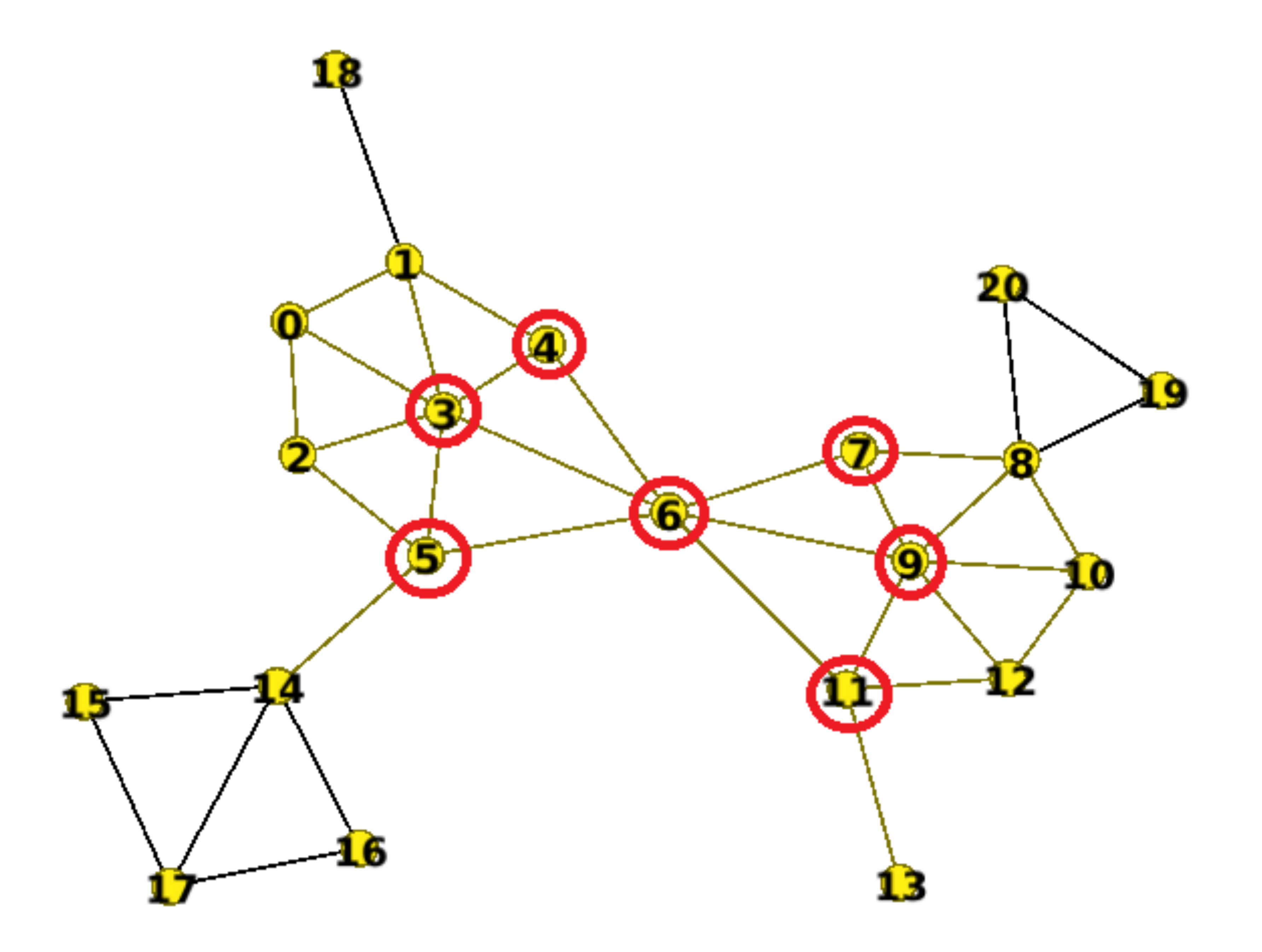


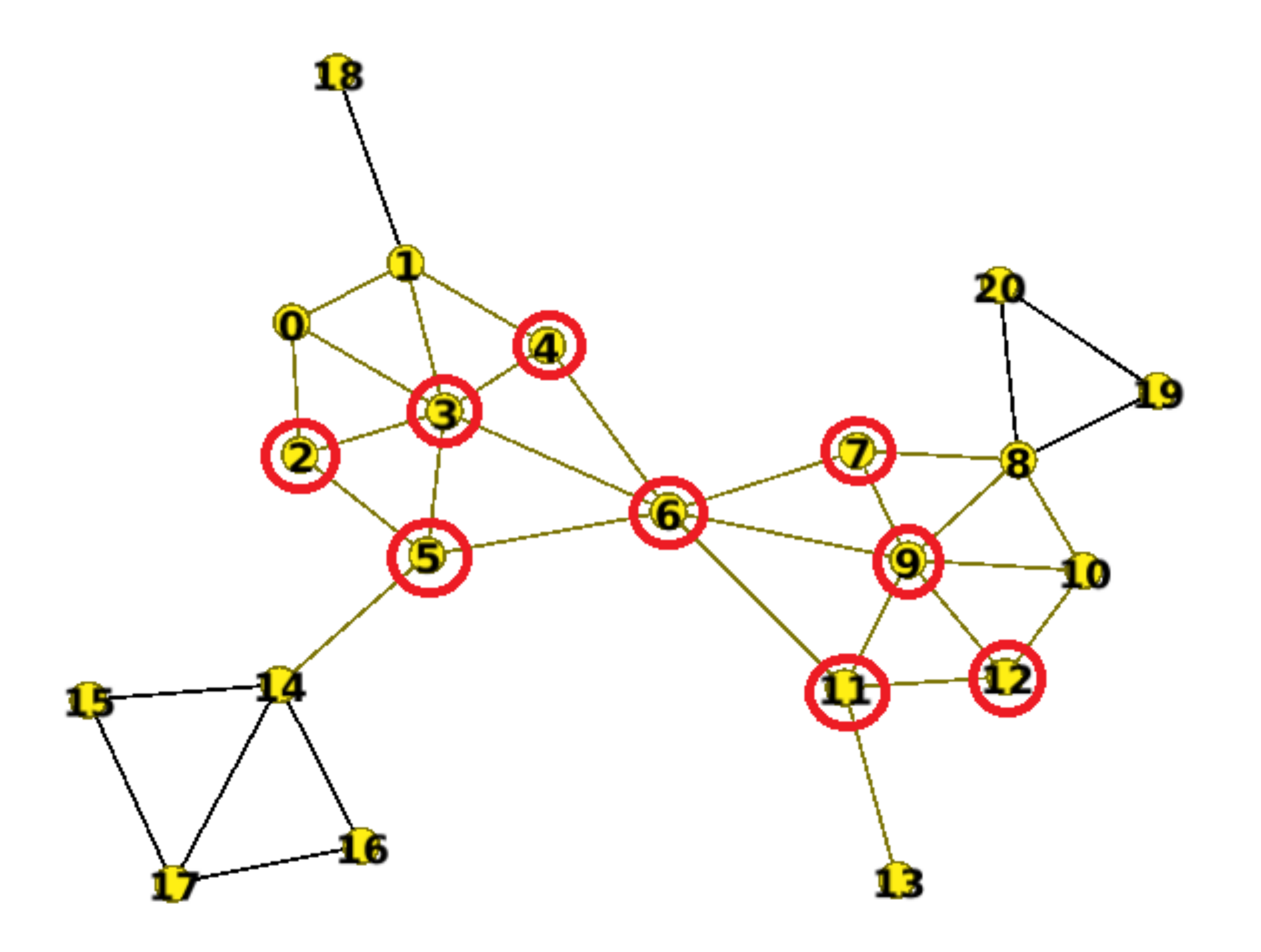
Communautés locales recouvrantes

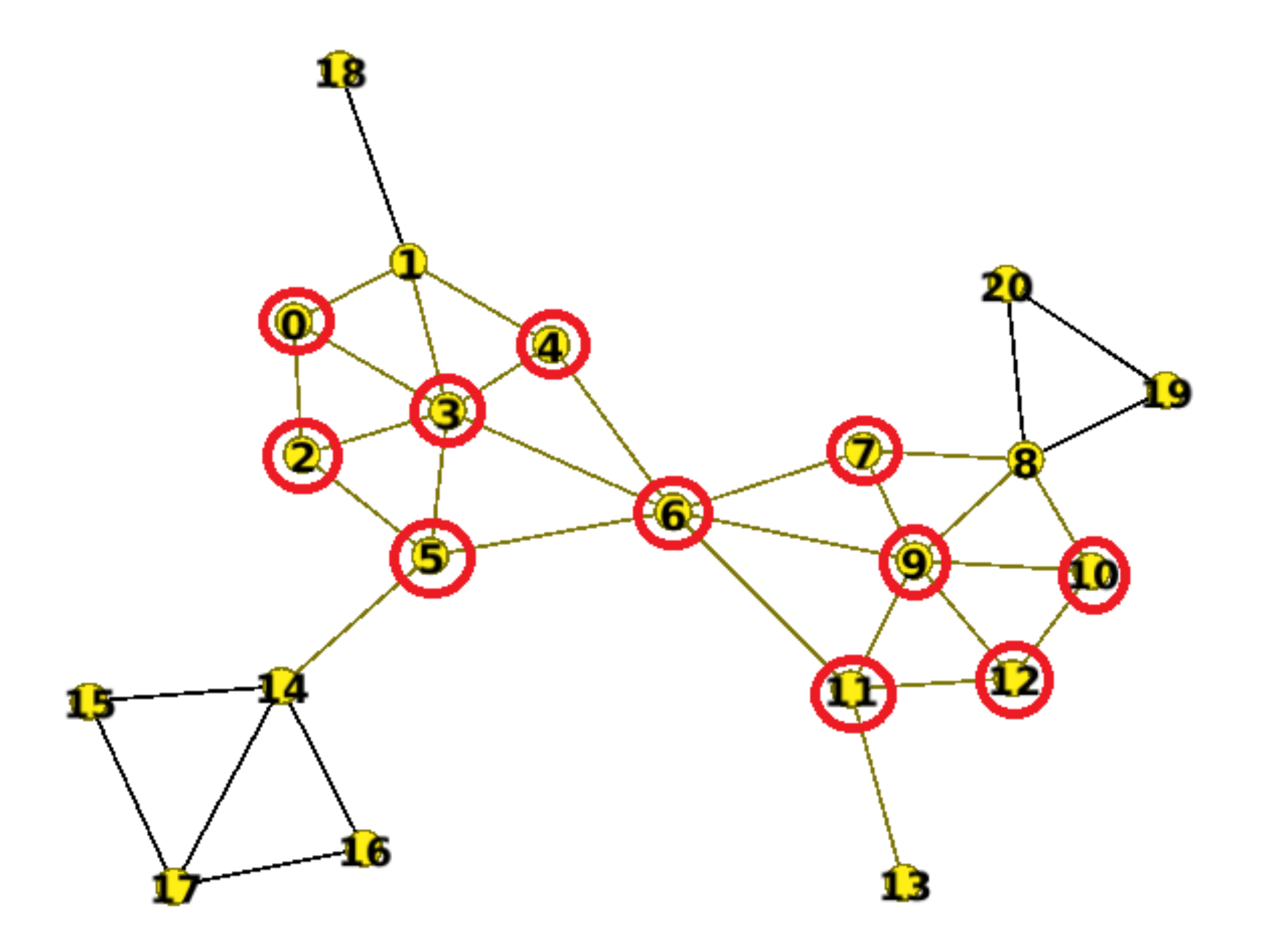


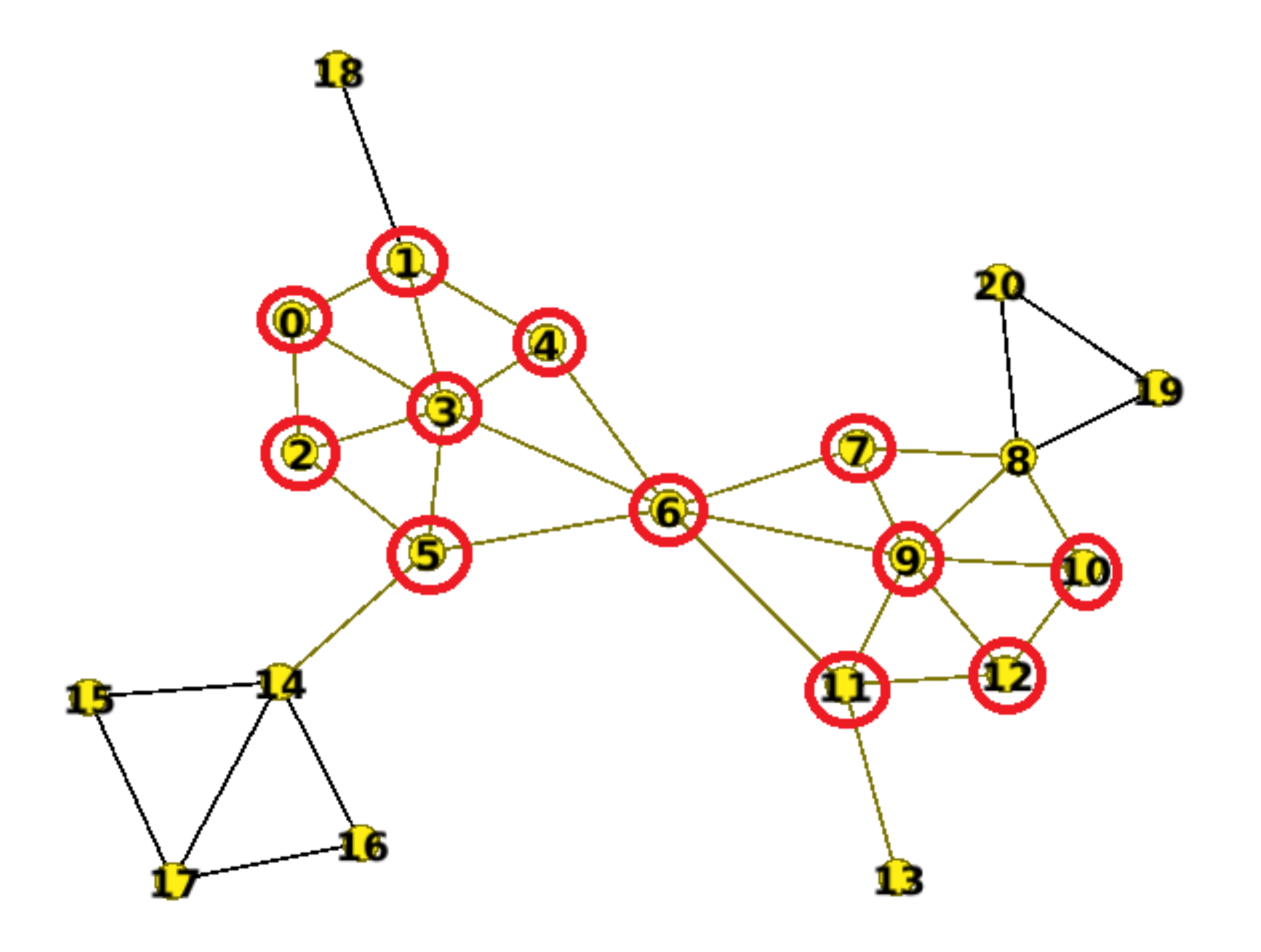

18
1
0
4
3
2
6
5
7
20
19
8
9
10
11
12
13
14
15
16
17

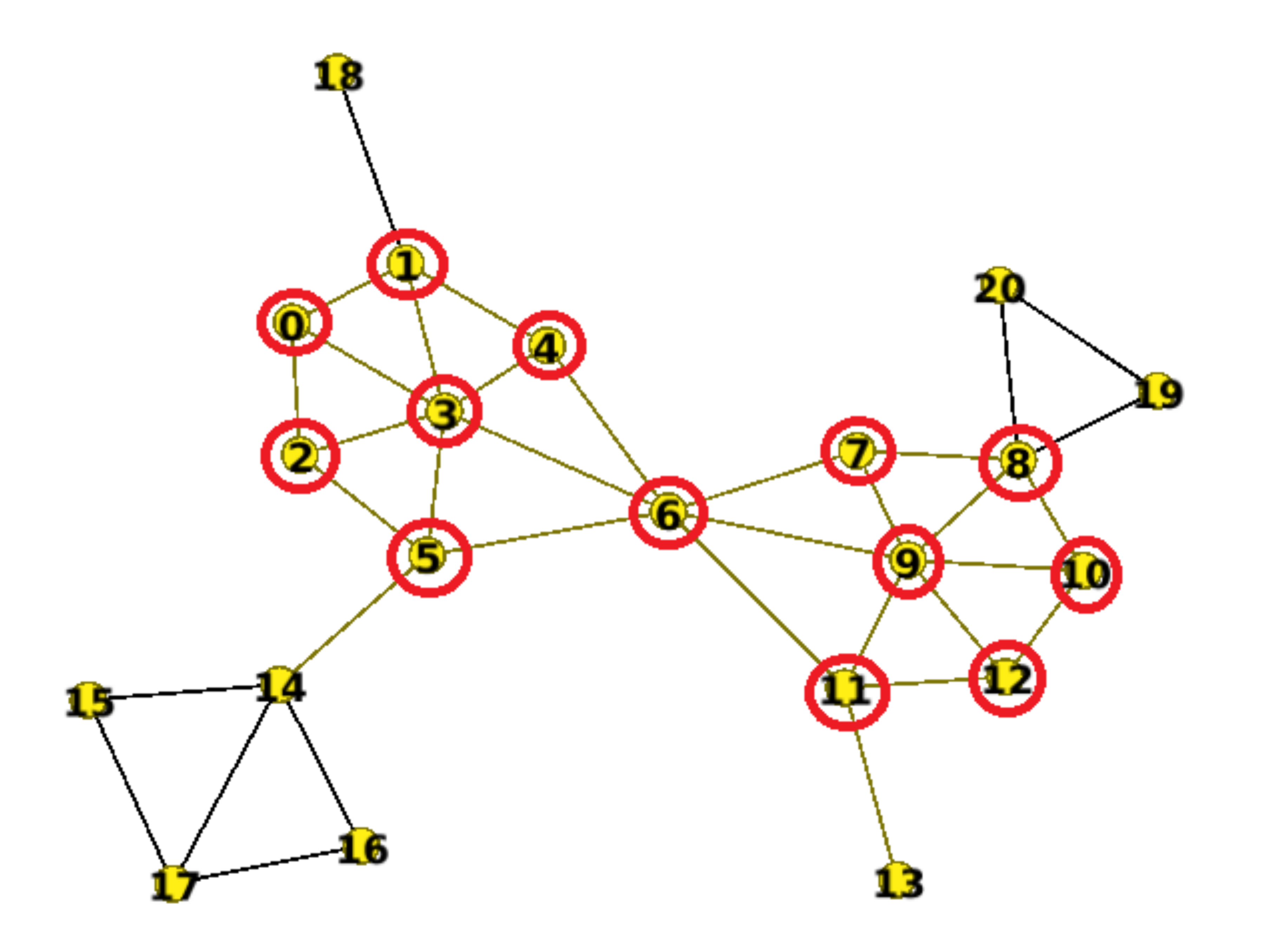

18
1
0
4
3
2
5
6
7
9
8
20
19
10
11
12
13
14
15
16
17

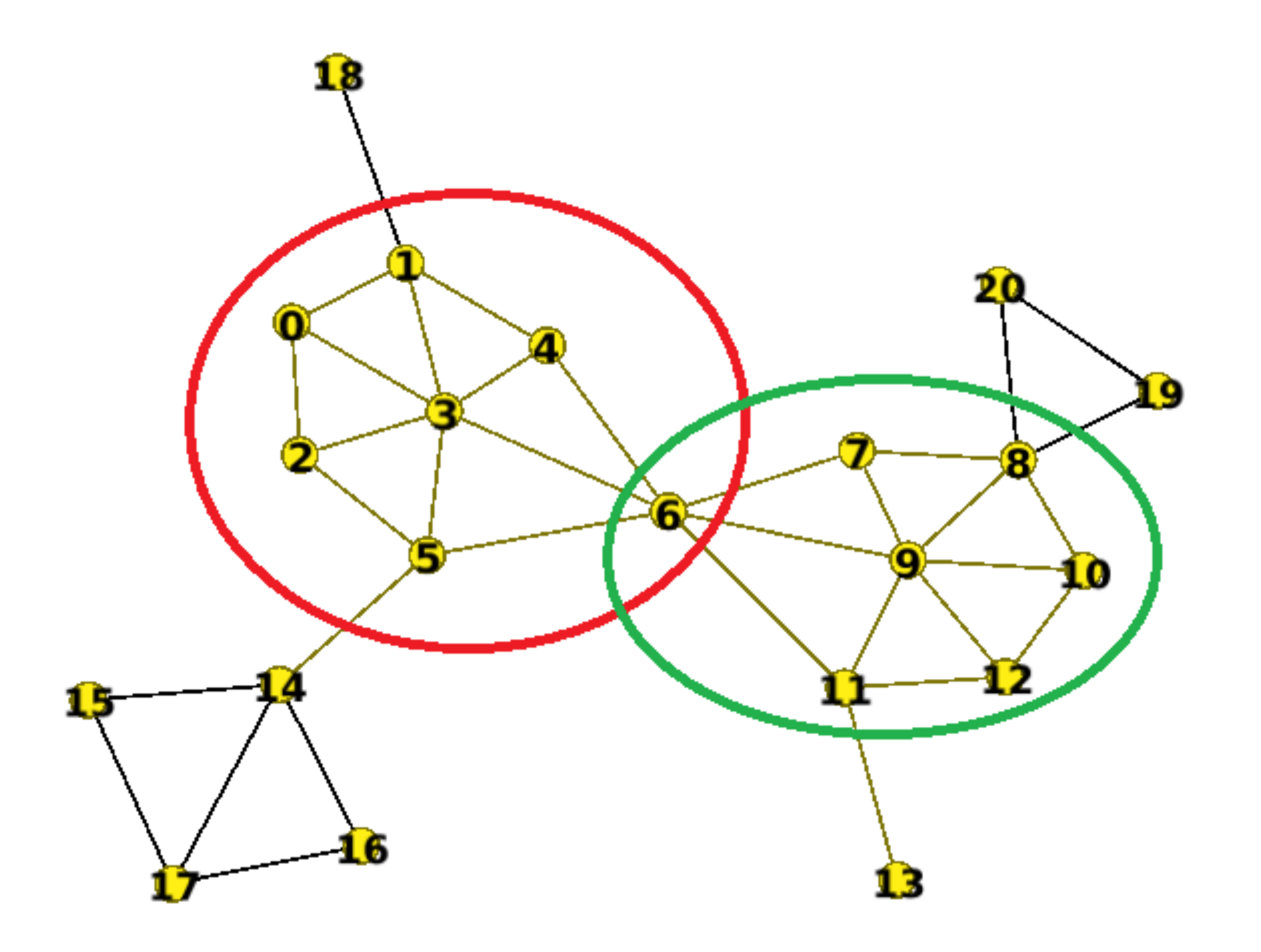

18
1
0
4
3
2
5
6
7
9
8
20
19
10
11
12
13
14
15
16
17


18
1
0
4
3
2
6
5
7
9
8
20
19
10
11
12
13
14
15
16
17


18
1
0
4
3
2
5
6
7
8
20
19
9
10
11
12
13
14
15
16
17


18
1
0
4
20
3
19
2
7
8
6
9
10
5
11
12
14
15
16
17
13


18
1
0
4
3
2
5
6
20
19
7
8
9
10
11
12
13
14
15
16
17



18
1
0
4
3
2
5
6
7
8
20
19
9
10
11
12
13
14
15
16
17

Schéma général des algorithmes locaux:

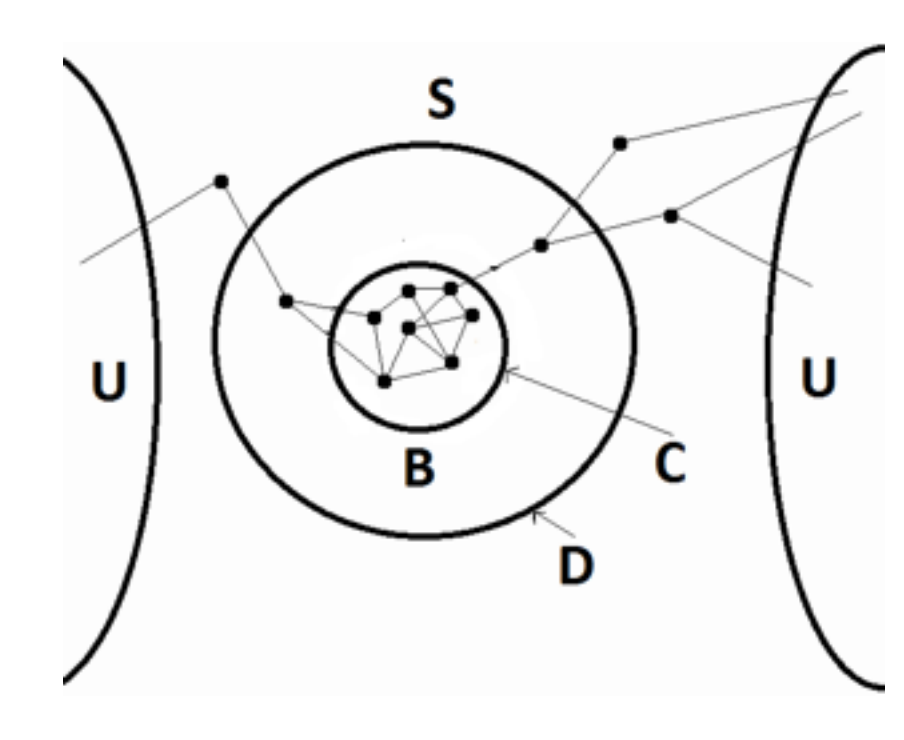
Algorithm 1 Identification of local communities based on greedy maximization of a quality criterion Q.

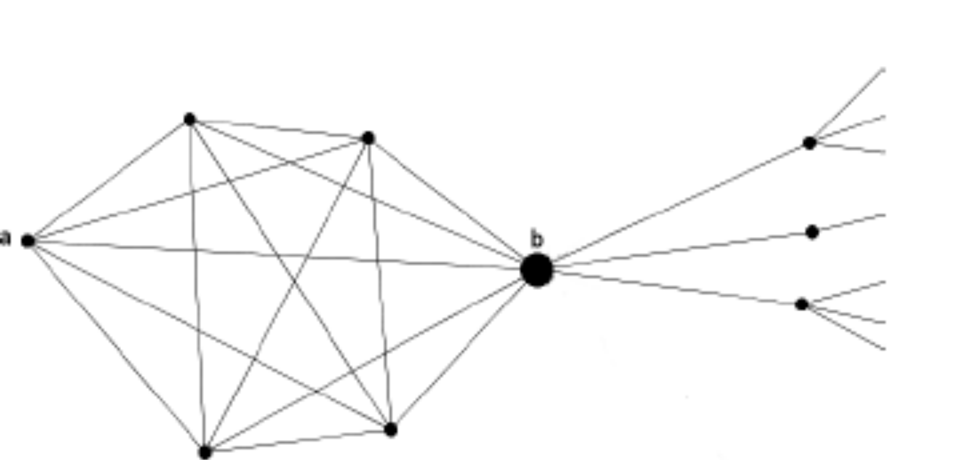
Algorithm: *Local community identification*

Input: a graph G and a starting node n_0.
Output: a subset D : the local community of n_0.
Initialize D with n_0
Initialize B with n_0
Initialize C with the empty set
Initialize S with the neighbors of n_0
Q = 0
Repeat
  For each $s_i \in S$ **do**
    Compute the quality criterion obtained if s_i is added to D
  End for each
  Select the node s^* that produces the maximal quality Q^*, breaking ties randomly.
  If $Q^* > Q$ **then**
    Add s^* to D and remove it from S.
    Update B, S, C.
  End if
Until $(Q^* \leq Q)$
Return D

Fonctions de «qualité»

Clauset: les nœuds de *B* doivent avoir plus de liens avec *C* qu'avec *S*





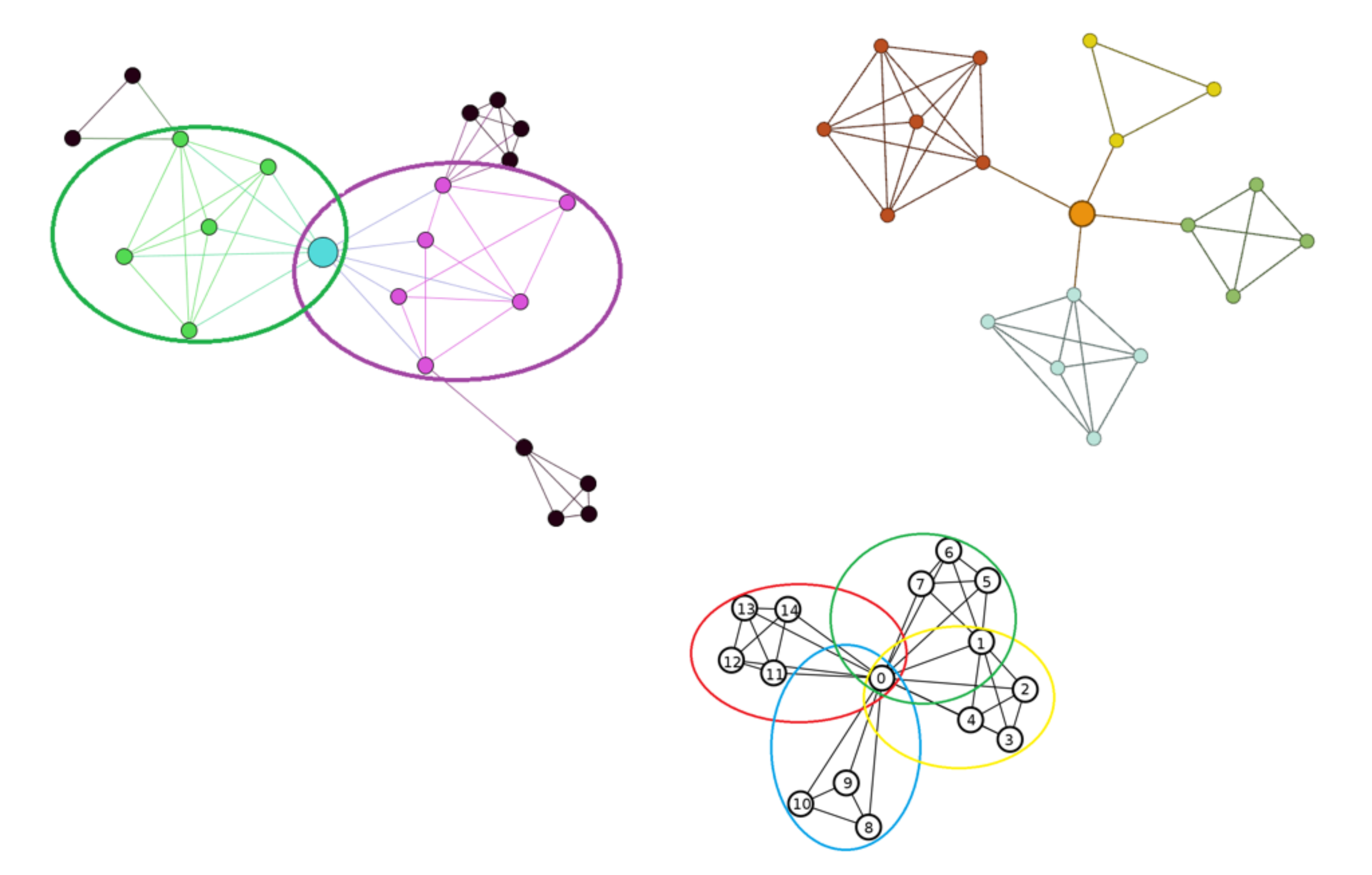
$$R = \frac{B_{in}}{B_{in} + B_{out}}$$

Raffinements par Luo, puis Chen.

Notre approche:

- prendre aussi en compte la distance au nœud de départ
- deuxième passe pour détecter sous-communautés recouvrantes (percolation de cliques)

Communautés locales recouvrantes:



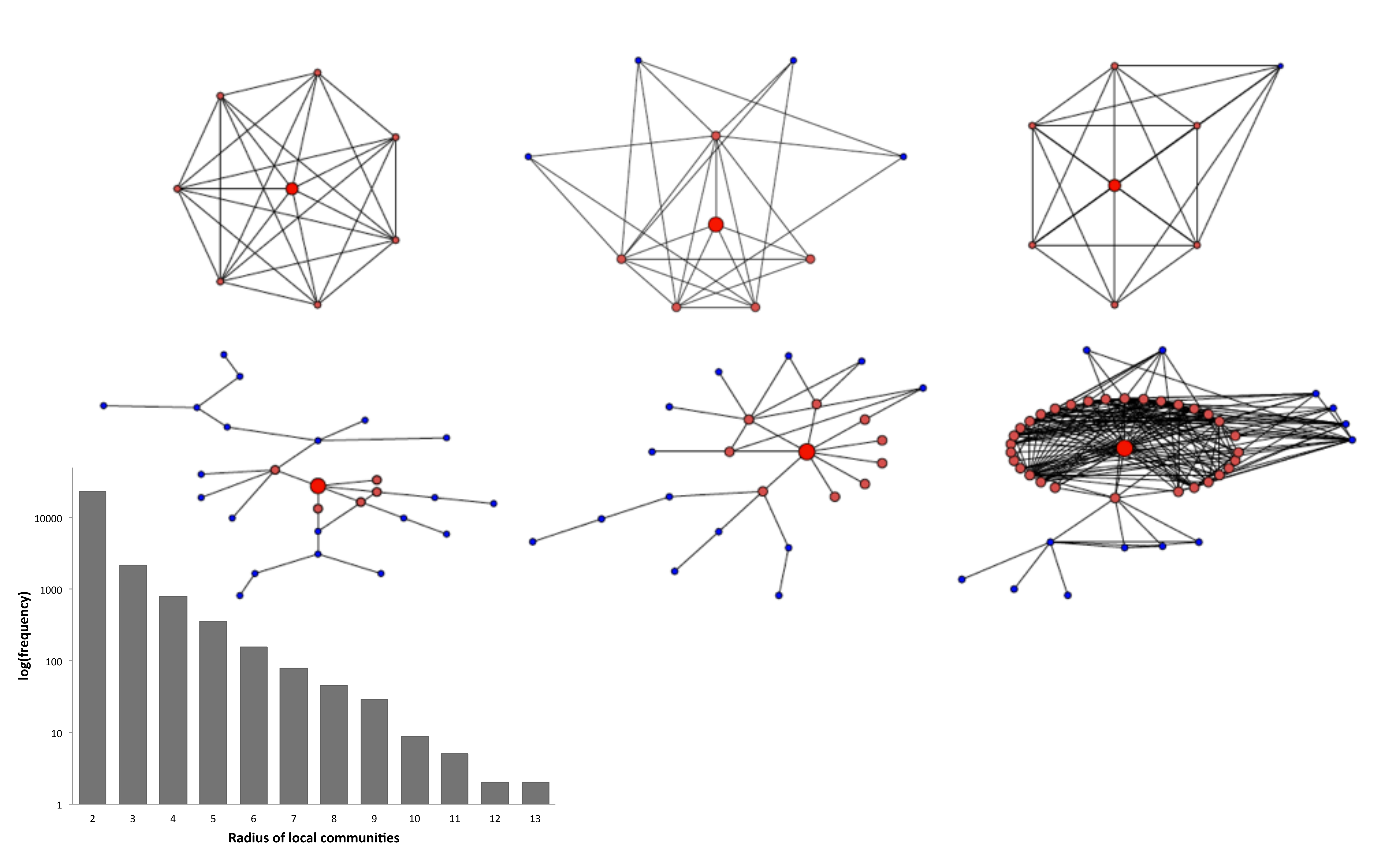
Comparaison avec l'algorithme de référence (Chen)

- Jeux de données NCAA Football 2000.
- 115 équipes organisées en conférences
- Structures de communautés connue : les conférences.

2000 NCAA Football	Résultats des algorithmes	
	Algorithme de Chen	Notre algorithme
Pas de Communautés	41(35,65%)	0(0%)
Précision	0,98	0,93
Rappel	0,82	0,90
F-Mesure	0,87	0,91

➡ Ngonmang, Tchuente, Viennet, *Parallel Processing Letters*, 2012

Exemples de communautés locales:



Distribution des degrés et tailles de communautés



1000000
100000
10000
1000
100
10
1

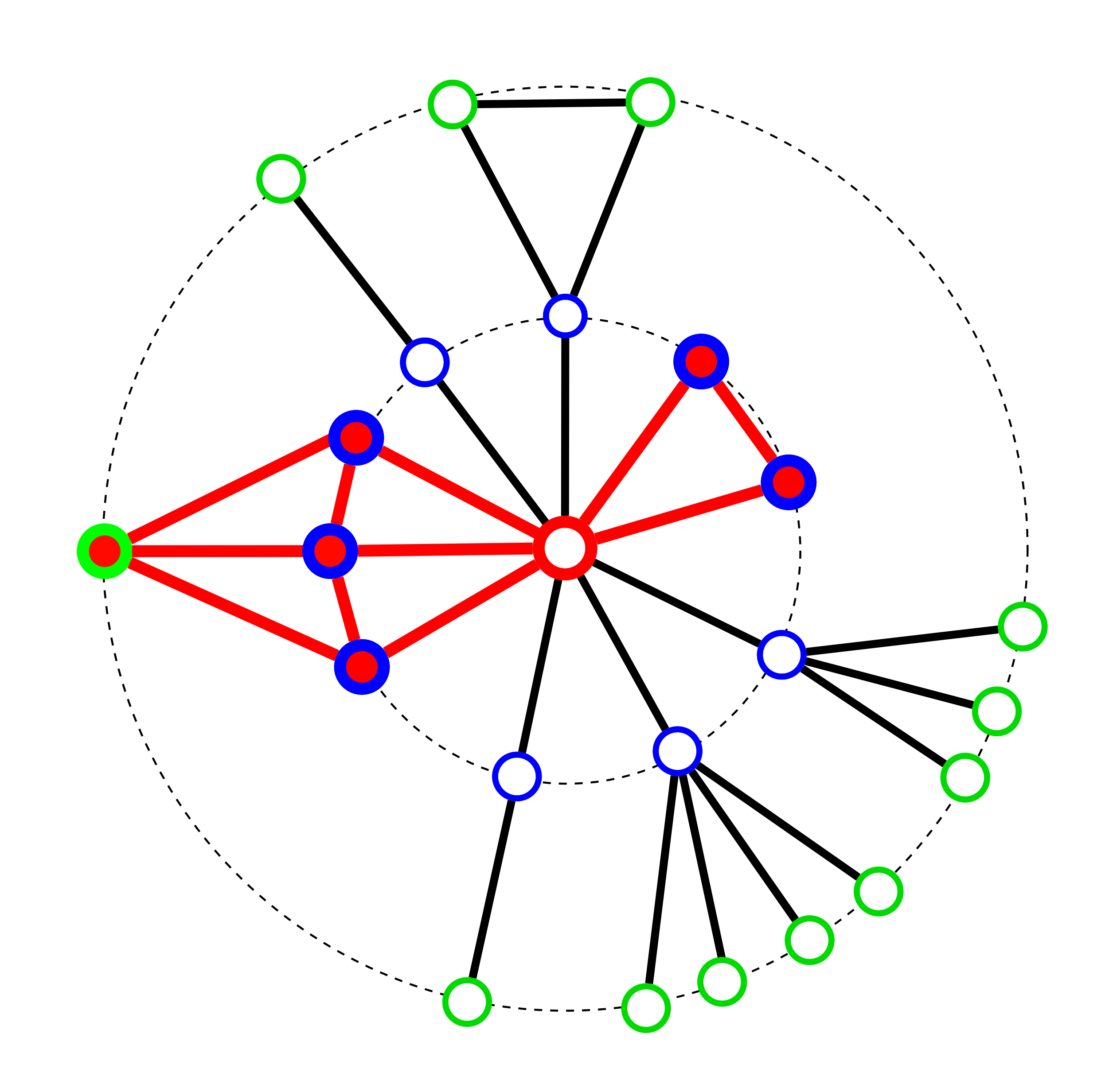
Frequency

1 10 100 1000 10000

Node degree (top) or Community size (bottom)

Application : prévision de *churn*





Périodes d'apprentissage et test pour le modèle de *churn*

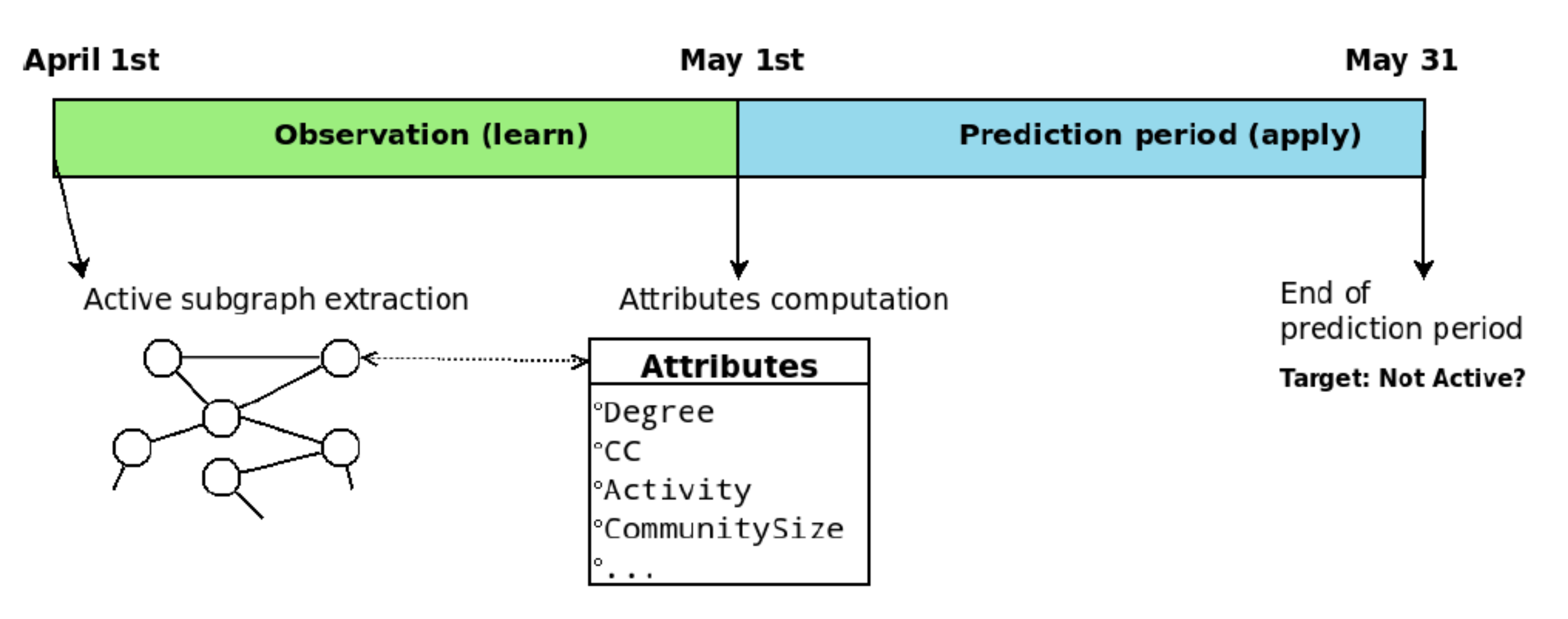


Fig. 4. Experimental setup used for churn prediction.

Indicateurs calculés sur chaque nœud:

Nœud lui même :

	Attribute name	Description
1	Degree	The degree of the node
2	CC	The local clustering coefficient of the node
3	Activity	The number of time the node has made a connexion
3	DaysAfterLastCon	The number of days since the last connexion of the node

Sur sa communauté locale :

4	LocalComSize	Size of the local community i.e. the number of nodes of the local community
5	LocalInProp	The internal proportion i.e. the proportion of local community's node directly connected to the starting node
6	LocalAvgDegree	The Average degree of the nodes inside the local community
7	LocalPropInact	The proportion of nodes inside the local community that are already inactive
8	LocalAvgAct	The average activity for the nodes of the local community

Indicateurs calculés sur chaque nœud (suite):

Sur le premier cercle :

9	NeigSize	Size of the first Neighborhood
10	NeigAvgDegree	The Average degree of the first Neighborhood
11	NeigPropInact.	The proportion of nodes inside the first Neighborhood that are already inactive
12	NeigAvgAct	The average activity for the nodes of the first Neighborhood

Sur le deuxième cercle :

13	Neig2Size	Size of the second Neighborhood
14	Neig2AvgDegree	The Average degree of the second Neighborhood
15	Neig2PropInact	The proportion of nodes inside the first Neighborhood that are already inactive
16	Neig2AvgAct	The average activity for the second Neighborhood

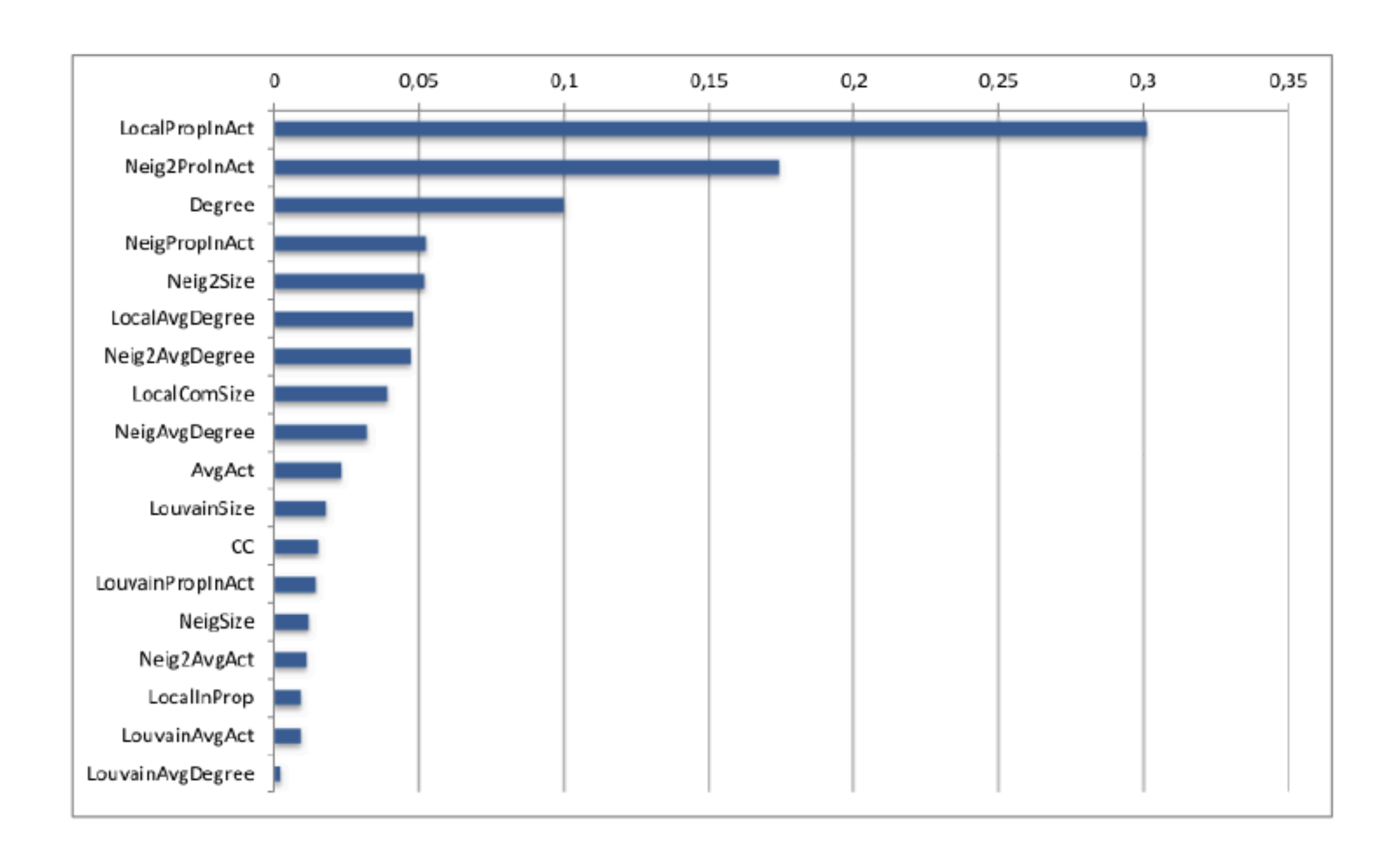
Sur sa communauté globale (Louvain):

17	LouvainSize	Size of the Louvain's global community the node belongs to
18	LouvainAvgDegree	The Average degree of the Louvain's global community the node belongs to
19	LouvainPropInact.	The proportion of nodes inside the Louvain's global community the node belongs to that are already inactive
20	LouvainAvgAct	The average activity for the Louvain's global community the node belongs to

Attributes sets	Avg #nodes used	Accuracy	Precision	Recall	F-Score	AUC
All	431978	79.3 %	64.0 %	73.5 %	0.684	0.855
All without Louvain's global community	72353	79.2%	63.7 %	74.0 %	0.685	0.854
Node & local community	21	79.0 %	63.2 %	74.8 %	0.686	0.832
Node & second Neighborhood	71734	78.9 %	63.0 %	75.2 %	0.686	0.826
Node & first Neighborhood	598	78.9 %	63.0 %	74.9 %	0.685	0.824
Node & Louvain's global community	359625	78.9%	63.1%	74.7%	0.684	0.823
Node only	1	78.8 %	62.6 %	75.7 %	0.685	0.815
Local community only	20	71.5 %	53.4%	53.5%	0.530	0.727
Second neighborhood only	71733	68.4 %	48.4 %	52.5%	0.504	0.699
First neighborhood only	598	65.6 %	44.2%	47.4%	0.457	0.649
Louvain community only	359624	55.2 %	37.4%	69.3%	0.486	0.635

(gaussian SVM classifier)

Contribution des variables (modèle KXEN K2C/K2R)



Quelques perspectives de l'équipe

- suivi de communautés (dynamique)
- prévision de liens (recommandation)
- implémentations distribuées (cluster, «cloud»)